Statistica Sinica **24** (2014), 1879-1899
doi:http://dx.doi.org/10.5705/ss.2013.026

FDR CONTROL IN MULTIPLE TESTING UNDER NON-NORMALITY

Bing-Yi Jing, Xin-Bing Kong and Wang Zhou

HKUST, Soochow University, National University of Singapore

Abstract: There is a growing literature on the large-scale multiple testing in which the Benjamini-Hochberg (BH) procedure and its variants play a key role. Almost all this work assumes that the underlying distribution is normal in calculating the p-values. Here we study the effect of non-normality on false discovery control in large-scale multiple testing. The normal approximation, bootstrap calibration and the skewness-corrected normal approximation methods of approximating the individual p-values used to rank the significance levels are investigated. As an illustration, we compare these procedures with the BH method in terms of the cutting threshold and the false discovery rate.

Key words and phrases: False discovery rate, multiple testing.

1. Introduction

There has been an increasing interest in recent years in multiple testing problems arising from genomic data analysis, neuroimaging, and technical trading performance in financial markets, see for example, Abbas et al. (2013), Liu and Shao (2013), Barras, Acaillet, and Wermers (2010), and Bajgrowicz and Scaillet (2012). In such areas, often tens of thousands of tests are performed simultaneously. In this paper, we focus on the multiple testing rule with a common threshold for all p-values so that some kind of compound error rate is under control. In the seminal work, Benjamini and Hochberg (1995) (BH) suggest using the false discovery rate (FDR) as the compound error rate, and a step-up adaptive procedure is proposed and proved to control the FDR. Table lists the result of a large-scale test. Mathematically, $FDR = E(V/(R \vee 1)) = E[FDP]$.

If the p-value corresponding to the ith individual test is P_i, the BH adaptive procedure aims to find

$$\kappa = \max\{i; P_{(i)} \leq \frac{i}{m}\alpha\}, \tag{1.1}$$

where α is the nominal FDR level, $P_{(i)}$'s are sorted p-values in ascending order. The rule is to reject all null hypotheses corresponding to $P_{(i)}$, $i = 1, \ldots, \kappa$.

The BH adaptive procedure results in the control of FDR under level α. A modified BH procedure is later developed to incorporate an estimate of the

Table 1. Number of errors committed when testing m hypotheses.

	Non-significant	Significant	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	$m - m_0$
	$m - R$	R	m

true null rate into consideration, see Benjamini and Hochberg (2000). Instead of fixing a FDR level first and then finding the cut point, Storey (2002) proposes a direct way that prescribes a cut threshold first and then estimates the FDR level. A natural conservative estimate of FDR, for a prescribed cutting threshold t, is

$$\widehat{FDR}(t) = \frac{m\hat{\pi}_0(\lambda)t}{R(t)}, \tag{1.2}$$

where $R(t)$ is the total number of rejections, $R(t) = \sharp\{i;\ P_i \le t,\ i = 1, \dots, m\}$, and $\hat{\pi}_0(\lambda) = \sharp\{i; P_i > \lambda\}/m(1-\lambda)$ is a conservative estimate of the true null rate π_0, where λ is a tuning parameter in $[0, 1)$. To gain more power, the optimal threshold is

$$t_\alpha(m) = \sup\{t : \widehat{FDR}(t) \le \alpha\}. \tag{1.3}$$

The multiple testing rule is to reject all hypotheses with $P_i \le t_\alpha(m)$.

When $\lambda = 0$ (or equivalently $\hat{\pi}_0(\lambda) \equiv 1$), the Storey procedure (1.3) is equivalent to the BH procedure (1.1). Procedures (1.1) and/or (1.3) are later developed and refined from either the perspective of FDR control or FDR estimation, see for example, Benjamini and Yekutieli (2001), Benjamini, Krieger, and Yekutieli (2006), Genovese and Wasserman (2004), Wu (2008), Hu, Zhao, and Zhou (2011), Nettleton et al. (2006), and Liang and Nettleton (2012).

All the work is based on the assumption that the p-values can be calculated accurately, while in practice the calculation of p-values is not easy and the null distribution of the individual test statistic has to be known. If the population distribution is normal, the t-statistic can be used for testing means, But when the population distribution departs from normality, one has to approximate the distribution of t-statistics using the normal or the bootstrap-t distribution. The problem here is that the sample size is typically small- in microarray studies, for example, the sample size is usually in the tens, while the dimensionality is very large. This results in the inaccuracy of individual p-values, and the effect of this inaccuracy on the control of FDR is not well understood.

This problem fits into the small-n-large-p paradigm. In this paper, we give expansions of $t_\alpha(m)$ in terms of approximate cutting thresholds, and expansions of the false discovery rate in terms of approximate p-values. While the heavy-tailed phenomenon does not affect the approximate cutting thresholds and the

false discovery rate much, the asymmetry (skewness) of the population distribution does. We focus on the BH procedure. Investigation into the effect of the inaccurate individual p-values on the estimation of π_0 and the choice of λ in $\hat{\pi}_0(\lambda)$ is left as a separate project.

As to the effect of the normal approximation (NA) on the p-values, we show that the approximate cutting threshold deviates to the left/right side of the oracle threshold depending on the sign of the skewness, resulting in too conservative/liberal control of false discovery rate. For the bootstrap-t calibration method on the p-values, We show that the approximate cutting threshold has second order accuracy to the oracle cutting threshold leading to much better control of false discovery rate. We propose a skewness-corrected normal approximation (SC) method when the sign of the skewness is consistent with the direction of the test hypotheses in some sense. Instead of correcting for skewness using only the data within genes as in the bootstrap method, our method corrects for skewness using the data within and across genes leading to better efficiency. Numerical examples confirm our findings.

Related to this work, Fan, Hall, and Yao (2007) study the effect of the NA and bootstrap-t calibration on the individual p-value calculation and on the control of the family wise error rate using Bonferroni's approach. Delaigle, Hall, and Jin (2011) reveal the robustness and accuracy of the bootstrap-t calibration in the individual p-value calculation with applications to the higher criticism test based on the bootstrap-t distribution for detecting sparse signals. To the best of our knowledge, there is no literature investigating the effect of non-normality on the control/estimation of false discovery rate.

The present paper is organized as follows. In Section 2, we study the effect of the inaccuracy of p-values on the cut threshold and false discovery rate using three approximation methods. Simulations are reported on in Section 3. Section 4 presents a data example. Section 5 concludes. All proofs are postponed to the Appendix.

2. The Model and Approximate Threshold

2.1. Model

Consider the model

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad 1 \le i \le m, \quad 1 \le j \le n, \tag{2.1}$$

where for fixed i, the X_{ij}'s are random variables with mean μ_i and error ϵ_{ij}'s. In a gene micro-array study, X_{ij} denotes the expression level of gene i of array j, the μ_i's and ϵ_{ij}'s represent random mean effects and measurement errors of gene expressions, respectively. In finance, X_{ij} models the excess return of the

ith asset over the jth investing period, μ_i and ϵ_{ij} represent the expected excess return and the risk including the market risk and idiosyncratic risk, respectively. We need a technical assumption.

Assumption 1.

1. *The μ_i's are i.i.d. continuous random variables with a mixture distribution of a point mass δ_0 at the origin and a distribution $\mu(x)$ with no point mass at the origin.*
2. *The ϵ_{ij}'s, $i = 1, \ldots, m$, $j = 1, \ldots, n$, are i.i.d. centered continuous random variables; jointly the $(\mu_i, \epsilon_{i1}, \ldots, \epsilon_{in})$, $i = 1, \ldots, m$, are i.i.d. continuous random vectors.*

In gene micro-array studies, it is of interest to find the over- and/or under-expressed genes; in finance, one hopes to identify the well-handled and/or badly managed mutual funds, c.f., Barras, Acaillet, and Wermers (2010) and Bajgrowicz and Scaillet (2012). This is formally equivalent to testing against $H_i : \mu_i = 0$, $i = 1, \ldots, m$. Then H_i serves as an index having Bernoulli distribution with $Pr(H_i = 0) = \pi_0$ and $Pr(H_i = 1) = \pi_1$, $H_i = 0$ representing the null is true. We assume that the alternative hypotheses are that $\mu_i > 0$, $i = 1, \ldots, m$, we are interested in finding the over-expressed genes or well-managed mutual funds. Part of our theory extends to two-sided tests and also to two-sample problems (for example, testing for differently expressed genes in case control studies) without much effort.

The t-statistic is widely used in testing means and known for its robustness to outliers. Here we use the t-statistic in order to calculate p-values. Let $T_i = \sqrt{n}\overline{X}_i/S_i$, where

$$\overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}, \quad \text{and} \quad S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2. \tag{2.2}$$

If the ϵ_{ij}'s are normally distributed r.v.s, the p-values can be calculated exactly, but when the ϵ_{ij}'s are skewed and/or heavy-tailed, we have to use approximate distributions of the T_i's to find them. The latter is common in many practical applications. For example, in the leukemia gene expression data studied in Section 4, the estimated skewness is as high as 1.86. Another example comes from finance, where it is common that the risk measured by the fluctuation of the return time series is skewed, yielding the mean-variance-skewness efficient portfolio analysis in recent developments, c.f., Briec, Kerstens, and Jokung (2007) and Sentana (2008).

Let $G_0(n,t)$, $G_1(n,t)$, and $F(n,t)$ be, respectively the null, alternative, marginal c.d.f. of T_i's. Let $f(n,t)$ be the marginal p.d.f. of the T_i's with $\overline{F}(n,t) = 1 - F(n,t)$. For simplicity we suppress the dependence on n, and

write the above notation as $G_0(t)$, $G_1(t)$, $F(t)$, $f(t)$ and $\overline{F}(t)$, while noting that any condition on these quantities implicitly assumes that it holds uniformly in n.

Assumption 2. *$f(n,t)$ is Lipschitz continuous uniformly in all n.*

To find individual p-values, we use the NA, bootstrap calibration, and SC to derive $G_0(t)$, then study the effect of the approximation on the false discovery control. Since we have to approximate the distributions of a large number of T_i's, we consider the asymptotic regime as $n, m \to \infty$. We fix α, the nominal FDR level, across n, the sample size. When condsidering the finite sample performance, α is fixed at an appropriate location relative to the magnitude of n.

2.2. Normal approximation

Let P_i^N be the approximate p-value using the NA and $\tilde{T}_i$ be the observed t-statistic. $P_i^N = 1 - \Phi(\tilde{T}_i)$ with p.d.f. denoted by $w^N(t)$. The *normally approximated threshold* in testing m hypotheses simultaneously is

$$t_\alpha^N(m) = \sup\{t; \frac{mt}{\sharp\{i;\ P_i^N \le t\}} \le \alpha\}. \tag{2.3}$$

Take the *estimated FDR* (unknown) by the BH procedure committed at $t_\alpha^N(m)$ be

$$\widehat{FDR}(t_\alpha^N(m)) = \frac{mt_\alpha^N(m)}{\sharp\{i; P_i \le t_\alpha^N(m)\}}. \tag{2.4}$$

Similarly, in the limiting sense, take the counterpart cutting thresholds

$$t_\alpha^N = \sup\{t; H^N(t) := \frac{t}{Pr(P_1^N \le t)} \le \alpha\},\ t_\alpha = \sup\{t; H(t) := \frac{t}{Pr(P_1 \le t)} \le \alpha\}. \tag{2.5}$$

The $t_\alpha(m)$ as (1.3) is known if the P_i's are known, we call it the *oracle threshold*. Typically, the larger the cutting threshold, the more power the procedure has but the larger the FDR level. Thus underestimation of the oracle threshold brings a lower FDR but also less power, and overestimation implies more power but the FDR might not be well controlled. This is made precise by the following theorem.

Theorem 1. *Under Assumptions* 1 *and* 2*, suppose $E\epsilon_{11}^4$ is bounded above by a constant B, and $\Phi^{-1}(1 - t_\alpha^N) = o(n^{1/4})$ for a prescribed $\alpha > 0$. Suppose $H'(t)$ is bounded below for $t \in (t_\alpha, t_\alpha^N)$ and that $w^N(t_\alpha^N) < \alpha^{-1} < w^N(0)$. Then*

$$\begin{aligned} &t_\alpha(m) - t_\alpha^N(m) \\ &= C_1(\alpha, t_\alpha^N, G_0^{-1}(1 - t_\alpha^N))(\Phi^{-1}(1 - t_\alpha^N) - G_0^{-1}(1 - t_\alpha^N))(1 + o_n(1)) + O_p(\frac{1}{\sqrt{m}}), \end{aligned}$$

where

$$C_1(\alpha,t,u)=\frac{\alpha f(\Phi^{-1}(1-t))g_0(u)\overline{F}(u)}{g_0(u)\overline{F}(u)-tf(u)}$$

and, if $0\le\Phi^{-1}(v)\le Bn^{1/4}$,

$$\Phi^{-1}(v)-G_0^{-1}(v)=\frac{1}{3}\gamma n^{-1/2}(\Phi^{-1}(v))^2-C_2(B)\{n^{-1/2}+(1+\Phi^{-1}(v))^3n^{-1}\},$$

where γ *is the skewness of* ϵ_{11} *and* $C_2(B)$ *is a positive constant depending only on* B.

Remark 1. The condition $w^N(t_\alpha^N)<\alpha^{-1}<w^N(0)$ is used to guarantee that $t_\alpha^N(m)=t_\alpha^N+O_p(1/\sqrt{m})$ and $t_\alpha(m)=t_\alpha+O_p(1/\sqrt{m})$, see Lemma A.1 in the Appendix.

Remark 2. Since α is fixed, the condition $0\le\Phi^{-1}(1-t_\alpha^N)\le Bn^{1/4}$ is automatically satisfied if n is large enough. If α is set such that t_α^N is small, then $\Phi^{-1}(1-t_\alpha^N)$ is typically large relative to the magnitude of $O(n^{1/4})$ when we consider the finite sample performance. In this case, the principal bias is caused by the presence of skewness γ. If the c.d.f of P_1^N is concave, often assumed in the literature, $C_1(\alpha,t,u)$ is positive uniformly for all α, t and u. Thus one can find that the *normally approximated threshold* underestimates the oracle threshold if the population distribution is positively skewed. Similarly the negative skewness leads to overestimation.

When using $t_\alpha^N(m)$, targeted as α, the *actual estimated FDR* satisfies the following decomposition.

Theorem 2. *Under the conditions of Theorem* 1,

$$\widehat{FDR}(t_\alpha^N(m))=\alpha-\alpha/3\gamma n^{-1/2}C_3(\Phi^{-1}(1-t_\alpha^N))(1+o_n(1))+O(n^{-1/2})+O_p(\frac{1}{\sqrt{m}}) \tag{2.6}$$

for any fixed α *satisfying* $0\le\Phi^{-1}(1-t_\alpha^N)\le Bn^{1/4}$ *, where* $C_3(u)=\frac{u^2f(u)}{\overline{F}(G_0^{-1}(1-t_\alpha))}$.

Remark 3. From Theorem 2, the *actual estimated FDR* biases downward if the population distribution is positively skewed and upward if it is negatively skewed consistent with the observation in Remark 2.

We next study the false discovery proportion that results from using the P_i^N's,

$$FDP_m^N(t)=:\frac{\sharp\{H_i=0,P_i^N\le t\}}{\sharp\{P_i^N\le t\}}=\frac{\sharp\{H_i=0,1-\Phi(\tilde{T}_i)\le t\}}{\sharp\{1-\Phi(\tilde{T}_i)\le t\}}. \tag{2.7}$$

The point limit, $FDR^N(t)$, is expected to be

$$FDR^N(t) =: \frac{\pi_0(1 - G_0(\Phi^{-1}(1-t)))}{Pr(P_1^N \le t)}. \tag{2.8}$$

Theorem 3. *Under the conditions in Theorem* 1,
$FDP_m^N(t_\alpha^N(m)) = \pi_0\alpha + \pi_0\alpha[\exp(-1/3n^{-1/2}(\Phi^{-1}(1-t_\alpha^N))^3\gamma)R - 1] + O_p(\frac{1}{\sqrt{m}})$,

where $R = 1+\theta(n,\alpha)(1+\Phi^{-1}(1-t_\alpha^N))n^{-1/2}+(1+\Phi^{-1}(1-t_\alpha^N))^4n^{-1}$, *and* $\theta(n,\alpha)$ *is a bounded constant for* α *satisfying* $0 \le \Phi^{-1}(1-t_\alpha^N) \le Bn^{1/4}$.

Remark 4. Theorem 3 shows that, in the limiting sense, the NA is too conservative while less powerful in detecting the signals when $\gamma > 0$. When $\gamma < 0$, the NA can be too liberal and result in the uncontrolled false discovery proportion. See the results in our simulation studies.

2.3. Bootstrap-t approximation

Let $X_{i1}^*, \ldots, X_{in}^*$ be a bootstrap resample from $\mathcal{X}_i = \{X_{i1}, \ldots, X_{in}\}$. Put $T_i^* = (1/\sqrt{n})\sum_{j=1}^n (X_{ij}^* - \overline{X}_i)/S^*$, where

$$S^{*2} = (n-1)^{-1}\sum_{j=1}^n (X_{ij}^* - \overline{X}_i)^2.$$

Denote the discrete distribution of T_i^* given $\mathcal{X}_i$ by G_b^i. Let $G_b := G_b^1$. We have $P_i^b = 1 - G_b^i(\tilde{T}_i)$, where P_i^b is the bootstrap p-value related to gene i with p.d.f. denoted by $w^b(t)$. If we substitute the P_i^b's into the multiple testing procedure (1.3), we end up with the *bootstrap approximated threshold* and the *actual estimated FDR* level (unknown)

$$t_\alpha^b(m) = \sup\{t; \frac{mt}{\sharp\{i;\ P_i^b \le t\}} \le \alpha\}, \quad \widehat{FDR}(t_\alpha^b(m)) = \frac{mt_\alpha^b(m)}{\sharp\{i;\ P_i \le t_\alpha^b(m)\}}. \tag{2.9}$$

Similarly, we define

$$t_\alpha^b = \sup\{t; H^b(t) := \frac{t}{Pr(P_1^b \le t)} \le \alpha\}. \tag{2.10}$$

Theorem 4. *Under Assumptions* 1–2, *suppose* $E\epsilon_{11}^4 \le B$, *that* $H'(t)$ *is bounded below for* $t \in (t_\alpha, t_\alpha^b)$, *and that* $w^b(t_\alpha^b) < \alpha^{-1} < w^b(0)$. *Then*

$$t_\alpha(m) - t_\alpha^b(m) = C_1(\alpha, t_\alpha^b, \Phi^{-1}(1-t_\alpha^b))D(1-t_\alpha^b)(1+o_n(1)) + O_p(\frac{1}{\sqrt{m}}),$$

where
$D(v) = -\frac{1}{3}E(\gamma - \hat{\gamma}_n(1))n^{-1/2}(\Phi^{-1}(v))^2 + C_4(B)\{n^{-1/2} + (1+\Phi^{-1}(v))^3n^{-1}\}$,

and $\hat{\gamma}_n(1)$ *is the sample skewness based on observations for gene* 1 *with divisor* n, *and* $C_4(B)$ *is a bounded positive constant for* v *satisfying* $0 \le \Phi^{-1}(v) \le Bn^{1/4}$.

Remark 5. The bias term due to skewness is of smaller order here than in Theorem 1 since $\hat{\gamma}_n(1)-\gamma = O_p(\sqrt{1/n})$. The bootstrap-$t$ approximation corrects the skewness automatically and is of higher order accuracy than the NA in the individual p-value calculation.

The *actual estimated FDR* satisfies the following decomposition.

Theorem 5. *Under the conditions in Theorem* 4,

$$\begin{aligned}&\widehat{FDR}(t^b_\alpha(m)) - \alpha \\ &\quad = \alpha/3E(\gamma - \hat{\gamma}_n(1))n^{-1/2}(\Phi^{-1}(1-t^b_\alpha))^2 C_3(G_0^{-1}(1-t^b_\alpha))(1+o_n(1)) \\ &\qquad +O(n^{-1/2}) + O_p(\frac{1}{\sqrt{m}}),\end{aligned}$$

for all α *satisfying* $0 \le \Phi^{-1}(1-t^b_\alpha) \le Bn^{1/4}$.

We next study the false discovery proportion that results from using the P_i^b's,

$$FDP^b_m(t) = \frac{\sharp\{H_i = 0, P_i^b \le t\}}{\sharp\{P_i^b \le t\}} = \frac{\sharp\{H_i = 0, 1 - G_b^i(\tilde{T}_i) \le t\}}{\sharp\{1 - G_b^i(\tilde{T}_i) \le t\}}. \tag{2.11}$$

Its limit, $FDR^b(t)$, is expected to be

$$FDR^b(t) = \frac{\pi_0 E(1 - G_0(G_b^{-1}(1-t)))}{Pr(P_1^b \le t)}. \tag{2.12}$$

Theorem 6. *Under the conditions in Theorem* 4, *if* $E\epsilon_{11}^8 \le B$ $(B > 1)$, *then*

$$\begin{aligned}FDP^b_m(t^b_\alpha(m)) = \pi_0\alpha &\left(1 + O[(1+\Phi^{-1}(1-t^b_\alpha))n^{-1/2} + (1+\Phi^{-1}(1-t^b_\alpha))^4 n^{-1}]\right) \\ &+O(n^{-1/2}) + O_p(\frac{1}{\sqrt{m}}),\end{aligned}$$

for all α *satisfying* $0 \le \Phi^{-1}(1-t^b_\alpha) \le Bn^{1/4}$.

Remark 6. Theorems 5–6 demonstrate that the bootstrap-t approximation is more robust in control of false discovery rate/proportion than the NA. To guarantee the $O(n^{-1/2})$ term, we need the eighth moment of ϵ_{11} in Theorem 6.

2.4. Skewness-corrected normal approximation

While the bootstrap calibration automatically corrects the skewness term, under the model (2.1) the skewness of X_{ij}'s under the null hypotheses is almost known since a very accurate estimate based on all $m \times n$ observations can be

used, see (2.13). Bootstrap calibration uses n observations to correct the skewness term, under (2.1), and loses efficiency in the estimation, resulting in less robustness and accuracy in approximating the *oracle threshold* and false discovery control. The computation is also time consuming when there are a lot of tests. Our estimator of skewness is

$$\hat{\gamma}_m = -\frac{\sum_{i=1}^{m}\sum_{j=1}^{[n/3]}(X_{i(3j)}+X_{i(3j-1)}-2X_{i(3j-2)})^3/([n/3]6m)}{\sum_{i=1}^{m}\sum_{j=1}^{[n/2]}(X_{i(2j)}-X_{i(2j-1)})^2/([n/2]2m)}. \tag{2.13}$$

By applying the Central Limit Theorem, we have the following.

Lemma 1. *Under Assumption* 1*, if* $E\epsilon_{11}^6 < \infty$*,*

$$\hat{\gamma}_m = \gamma + O_p(\frac{1}{\sqrt{m}}).$$

Let $1-\Phi_c(x) = (1-\Phi(x))\exp(-1/3n^{-1/2}x^3\hat{\gamma}_m)$. By Lemma 1, a slight variation of Theorem 4 of Delaigle, Hall, and Jin (2011), or Theorem 2 of Wang and Hall (2009), shows that under the null hypotheses,

$$\frac{\overline{G}_0(x)}{1-\Phi_c(x)} = 1+\theta(n,x)\{(1+|x|)n^{-1/2}+(1+|x|)^4n^{-1}\}+O_p(m^{-1/2}n^{1/4}) \tag{2.14}$$

as $n\to\infty$, where $\theta(n,x)$ is bounded by a finite positive constant $C_1(B)$ uniformly for all distribution of X with $E(|X/\sigma|^4)\le B$ $(B>1)$, and uniformly for all x satisfying $0\le x\le Bn^{1/4}$. When x is fixed the O_P term is of order $m^{-1/2}$.

We can calculate the p-value for gene i as $P_i^c = 1-\Phi_c(\tilde{T}_i)$ with p.d.f. denoted by $w^c(t)$, where $\tilde{T}_i$ is the observed t-statistic. Correspondingly, we define the *skewness-corrected normal approximated threshold* and the *actual estimated FDR* as

$$t_\alpha^c(m) = \sup\{t; \frac{mt}{\sharp\{i; P_i^c\le t\}}\le\alpha\},\ \widehat{FDR}(t_\alpha^c(m)) = \frac{mt_\alpha^c(m)}{\sharp\{i; P_i\le t_\alpha^c(m)\}}. \tag{2.15}$$

Similarly, we take

$$t_\alpha^c = \sup\{t; H^c(t) := \frac{t}{Pr(P_1^c\le t)}\le\alpha\}. \tag{2.16}$$

Theorem 7. *Under the conditions in Theorem* 1*, if* $w^c(t_\alpha^c) < \alpha^{-1} < w^c(0)$ *and* $\gamma>0$*, then*

$$\begin{aligned} t_\alpha(m)-t_\alpha^c(m) &= C_1(\alpha,t_\alpha^c,\Phi^{-1}(1-t_\alpha^c))C_2(B)\{n^{-1/2}+(1+\Phi^{-1}(1-t_\alpha^c))^3n^{-1}\}(1+o_n(1)) \\ &\quad +O_p(\frac{1}{\sqrt{m}}), \end{aligned}$$

for v *satisfying* $0\le\Phi^{-1}(v)\le Bn^{1/4}$.

Remark 7. We assume that $\gamma > 0$, otherwise $\Phi_c(x)$ is not well defined. In practice, this condition can be well identified and verified easily. Actually, the SC method works for negative skewness when we are testing for under-expressed genes.

Remark 8. From Theorem 7, the effect of the skewness is almost completely eliminated except for a term of order $O_p(1/\sqrt{m})$ which is typically small for large scale multiple testing. This renders a more robust and accurate approximation approach to finding the approximate threshold, compared with the NA and the bootstrap-t calibration.

The *actual estimated FDR* satisfies

Theorem 8. *Under the conditions in Theorem* 7*,*

$$\widehat{FDR}(t_\alpha^c(m)) = \alpha + O(n^{-1/2} + (1 + \Phi^{-1}(1 - t_\alpha^c))n^{-1}) + O_p(\frac{1}{\sqrt{m}}) \tag{2.17}$$

for α *satisfying* $0 \le \Phi^{-1}(1 - t_\alpha^c) \le Bn^{1/4}$.

We next study the false discovery proportion that results from using the P_i^c's defined as follows.

$$FDP_m^c(t) = \frac{\sharp\{H_i = 0, P_i^c \le t\}}{\sharp\{P_i^c \le t\}} = \frac{\sharp\{H_i = 0, 1 - \Phi_c(\tilde{T}_i) \le t\}}{\sharp\{1 - \Phi_c(\tilde{T}_i) \le t\}}. \tag{2.18}$$

Its limit, $FDR^c(t)$, is expected to be

$$FDR^c(t) = \frac{\pi_0(1 - G_0(\Phi_c^{-1}(1 - t)))}{Pr(P_1^c \le t)}. \tag{2.19}$$

Theorem 9. *Under the conditions in Theorem* 7*,*

$$FDP(t_\alpha^c(m)) = \pi_0\alpha\left(1 + O[(1 + \Phi^{-1}(1 - t_\alpha^c))n^{-1/2} + (1 + \Phi^{-1}(1 - t_\alpha^c))^4 n^{-1}]\right) + O_p(\frac{1}{\sqrt{m}}) \tag{2.20}$$

for α *satisfying* $0 \le \Phi^{-1}(1 - t_\alpha^c) \le Bn^{1/4}$.

Remark 9. From Theorems 8−9, the SC is more robust and accurate in the false discovery control than the two previous methods. This is seen in our numerical analysis.

3. Simulation

In this section, we conduct extensive simulations to check our findings. We generated data for $m = 6,000$ genes. For each gene, we generated $n = 30$

expression data. We set $\pi_0 = 0.9$ and let $\mu(x)$ be the c.d.f of $N(\mu_0, 0.05^2)$ with $\mu_0 = 0.05$, 0.10, 0.15, and 0.20. To produce skewness, we generated measurement errors from the following.

Model I Errors are $(\chi_1^2 - 1) * \sqrt{0.025/2}$, a right skewed centered r.v. with standard deviation 0.05, and signal to noise ratio ranging from 0.71 to 2.83;

Model II Errors are $\Gamma(3, 1/3) * \sqrt{0.025/2}$, a right skewed centered r.v. with standard deviation equal to 0.193, and signal to noise ratio ranging from 0.26 to 1.

For each gene, we calculated the t-statistics and then used the NA, bootstrap-t approximation and SC methods to find approximate p-values. For the bootstrap method, 1,000 resamples were drawn for each gene. The simulations were repeated 100 times, and the averaged approximate thresholds and false discovery proportions were recorded. The standard errors are given in parentheses, see Tables 2−3. We make the following remarks.

All FDP records are almost under the control of their nominal levels. But the bootstrap and SC methods outperform the NA in their accuracy to the nominal FDR levels. Across the board, the SC method works best. This is further confirmed by the boxplots in the right panel of Figure 1.

The average threshold obtained by the NA method is always less than those of the bootstrap and SC methods, consistent with Theorems 1, 4, and 7, see also the left panel in Figure 1. Correspondingly, the average FDP levels by the NA are always less than those of the bootstrap and SC methods, consistent with Theorems 3, 6, and 9.

Almost all average FDP levels are biased downward. One source of the bias is the fact that we conservatively estimate π_0 in this paper by 1, as in Benjamini and Hochberg (1995). The oracle upper bound of the false discovery rate according to Theorems 3, 6, and 9 is nearly $\pi_0\alpha$, equal to 0.09 and 0.135 respectively.

As the signal to noise ratio increases, the standard errors of the FDP records by all three methods decrease and the average thresholds have an increasing trend, expected since it is easier to detect the signals when they grow stronger.

For $\gamma < 0$, we conducted a similar simulation, but for generating the measurement error from $(1 - \chi_1^2) * \sqrt{0.025/2}$. The results are given in Table 4. We make the following remarks based on Table 4.

The NA method is too aggressive, resulting in large upward biases to the cutting threshold and the false discovery proportion. This is consistent with our theoretical findings. The bootstrap calibration method reduces the bias though it still leads to upward biases, due in part to the small sample size.

We also included the SC method although in Section 2.4 positive skewness is required. It appears that the approximate threshold and the false discovery

Table 2. The average approximate thresholds and false discovery proportions as well as their standard errors in the parentheses ($\gamma > 0$) for **Model I**. AT and FDP stand for the approximate threshold and false discovery proportion respectively. NA, BA and SC are respectively the short abbreviations of the NA, bootstrap approximation and skewness-corrected normal approximation.

μ		0.05	0.10	0.15	0.20
			$\alpha = 0.10$		
AT	NA	0.0033 (0.0003)	0.0074 (0.0004)	0.0096 (0.0004)	0.0101 (0.0004)
	BA	0.0049 (0.0003)	0.0087 (0.0004)	0.0103 (0.0004)	0.0107 (0.0004)
	SC	0.0051 (0.0004)	0.0090 (0.0005)	0.0105 (0.0004)	0.0110 (0.0004)
FDP	NA	0.0068 (0.0057)	0.0118 (0.0051)	0.0138 (0.0048)	0.0142 (0.0043)
	BA	0.0670 (0.0141)	0.0653 (0.0112)	0.0619 (0.0107)	0.0643 (0.0091)
	SC	0.0991 (0.0166)	0.0915 (0.0134)	0.0859 (0.0124)	0.0888 (0.0115)
			$\alpha = 0.15$		
AT	NA	0.0054 (0.0005)	0.0117 (0.0006)	0.0147 (0.0006)	0.0154 (0.0006)
	BA	0.0079 (0.0006)	0.0137 (0.0007)	0.0161 (0.0007)	0.0166 (0.0007)
	SC	0.0084 (0.0006)	0.0143 (0.0007)	0.0167 (0.0007)	0.0171 (0.0007)
FDP	NA	0.0146 (0.0078)	0.0229 (0.0080)	0.0275 (0.0071)	0.0283 (0.0067)
	BA	0.0974 (0.0167)	0.0944 (0.0134)	0.0968 (0.0144)	0.0992 (0.0122)
	SC	0.1374 (0.0186)	0.1272 (0.0143)	0.1283 (0.0146)	0.1280 (0.0132)

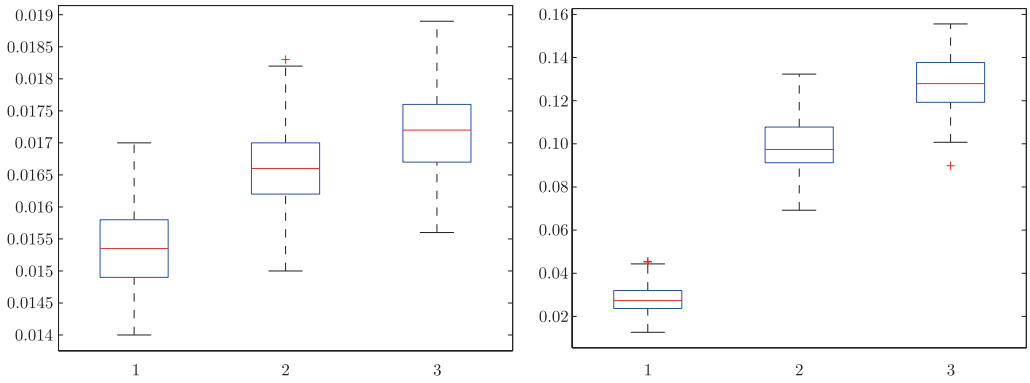

Figure 1. Boxplots of the approximated threshold (left panel) and FDP (right panel) for NA, BA, and SC. Here $\alpha = 0.15$, $\mu = 0.15$, and the data was generated from **Model I**.

proportion is biased downward severely across the board. The reason is that when $\gamma < 0$ and we observe a large t-statistics (this is typical in the simulation), $1 - \Phi_c$ renders an unreasonably large p-value, see (2.14). This leaves a potential signal undetected and hence a too-conservative procedure.

When $\gamma < 0$, the SC fails while the other two approaches are not satisfactory. In this case we are expecting a better approximation to make the FDP well

Table 3. The averaged approximate thresholds and false discovery proportions as well as their standard errors in the parentheses ($\gamma > 0$) for **Model II**. AT and FDP stand for the approximate threshold and false discovery proportion, respectively. NA, BA and SC are respectively the short abbreviations of NA, bootstrap approximation and skewness-corrected normal approximation.

μ		0.05	0.10	0.15	0.20
			$\alpha = 0.10$		
	NA	0.0070 (0.0004)	0.0098 (0.0004)	0.0105 (0.0004)	0.0106 (0.0004)
AT	BA	0.0075 (0.0004)	0.0102 (0.0005)	0.0109 (0.0003)	0.0110 (0.0004)
	SC	0.0078 (0.0004)	0.0105 (0.0004)	0.0112 (0.0004)	0.0113 (0.0005)
	NA	0.0521 (0.0112)	0.0528 (0.0108)	0.0549 (0.0090)	0.0534 (0.0096)
FDP	BA	0.0934 (0.0152)	0.0902 (0.0128)	0.0899 (0.0117)	0.0882 (0.0118)
	SC	0.1168 (0.0158)	0.1088 (0.0132)	0.1096 (0.0142)	0.1094 (0.0132)
			$\alpha = 0.15$		
	NA	0.0110 (0.0006)	0.0153 (0.0008)	0.0164 (0.0006)	0.0165 (0.0007)
AT	BA	0.0120 (0.0007)	0.0163 (0.0008)	0.0172 (0.0007)	0.0173 (0.0007)
	SC	0.0125 (0.0007)	0.0168 (0.0008)	0.0177 (0.0007)	0.0178 (0.0008)
	NA	0.0814 (0.0135)	0.0878 (0.0116)	0.0893 (0.0118)	0.0882 (0.0130)
FDP	BA	0.1341 (0.0177)	0.1341 (0.0155)	0.1314 (0.0138)	0.1299 (0.0140)
	SC	0.1620 (0.0176)	0.1607 (0.0157)	0.1563 (0.0153)	0.1540 (0.0143)

Table 4. The averaged approximate thresholds and false discovery proportions, with their standard errors in parentheses ($\gamma < 0$).

μ		0.05	0.10	0.15	0.20
			$\alpha = 0.10$		
	NA	0.0109 (0.0005)	0.0146 (0.0005)	0.0166 (0.0005)	0.0175 (0.0006)
AT	BA	0.0029 (0.0002)	0.0057 (0.0006)	0.0086 (0.0004)	0.0105 (0.0005)
	SC	0.0004 (0.0001)	0.0011 (0.0001)	0.0023 (0.0002)	0.0038 (0.0003)
	NA	0.5941 (0.0173)	0.4975 (0.0169)	0.4520 (0.0151)	0.4399 (0.0147)
FDP	BA	0.3903 (0.0369)	0.2737 (0.0231)	0.2211 (0.0190)	0.2035 (0.0162)
	SC	0.0983 (0.0558)	0.0330 (0.0202)	0.0157 (0.0125)	0.0103 (0.0058)
			$\alpha = 0.15$		
	NA	0.0189 (0.0008)	0.0247 (0.0010)	0.0278 (0.0009)	0.0175 (0.0006)
AT	BA	0.0052 (0.0006)	0.0102 (0.0008)	0.0149 (0.0009)	0.0105 (0.0005)
	SC	0.0006 (0.0001)	0.0016 (0.0002)	0.0035 (0.0003)	0.0038 (0.0003)
	NA	0.6231 (0.0180)	0.5366 (0.0174)	0.5025 (0.0148)	0.4399 (0.0147)
FDP	BA	0.4163 (0.0362)	0.3061 (0.0230)	0.2726 (0.0184)	0.2035 (0.0162)
	SC	0.0943 (0.0653)	0.0316 (0.0226)	0.0157 (0.0098)	0.0103 (0.0058)

controlled yet the BH procedure powerful.

Similar observations are illustrated by the boxplots in Figure 2.

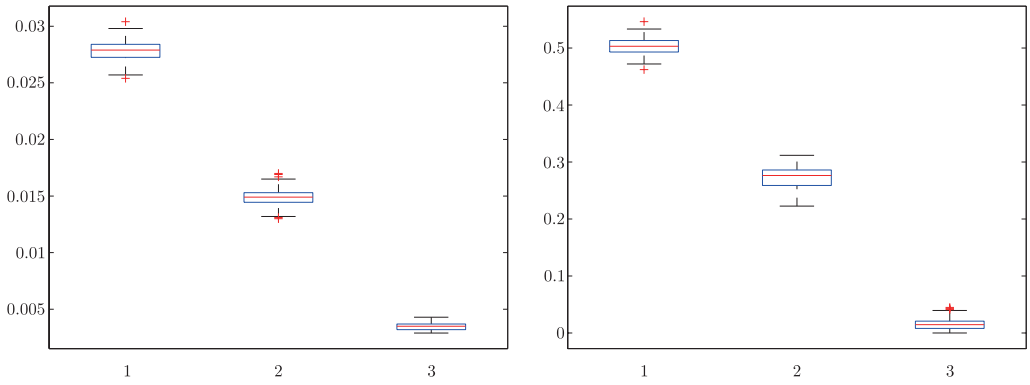

Figure 2. Boxplots of the approximated threshold (left panel) and FDP (right panel) for NA, BA, and SC. Here $\alpha = 0.15$, $\mu = 0.15$, and the measurement error was generated from $(1 - \chi_1^2) * \sqrt{0.025/2}$.

Table 5. The approximate thresholds using NA, BA, and SC under four nominal levels.

α		0.05	0.10	0.15	0.20
	NA	0.0105	0.0247	0.0418	0.0600
AT	BA	0.0107	0.0257	0.0442	0.0635
	SC	0.0128	0.0289	0.0463	0.0669

4. An Example

We applied our methods to the leukemia data of the expression levels of 5,000 genes from 27 patients. The data set is from Broad Institute and downloadable from `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`. One is interested in finding the over-expressed genes relative to the expression levels of AML patients, and this can be answered via multiple testing techniques. The raw data is adjusted by standardizing the expression levels of 27 observations for each gene and transforming the corresponding center to that of the AML expression levels. We then calculate the t-statistic for each gene. We deleted those genes with too large t-statistics in absolute value (> 5). Our ultimate data set consisted of 4,749 adjusted gene expression values and 4,749 t-statistics. Figure 3 shows the histogram of the centered expression data, which depicts a slight longer right tail than the left tail.

Our estimated skewness of the measurement error is 1.8625, far from the normal. From the perspective of FDR control, we have to find a common threshold to find interesting genes. Table 5 displays the estimated thresholds using the NA, BA, and SC methods for nominal levels $\alpha = 0.05$, 0.10, 0.15, and 0.20. Our observations are as follows.

The estimated thresholds increase from the first to the third row for fixed α, as expected when $\gamma > 0$. Our proposed SC method is the most powerful for all

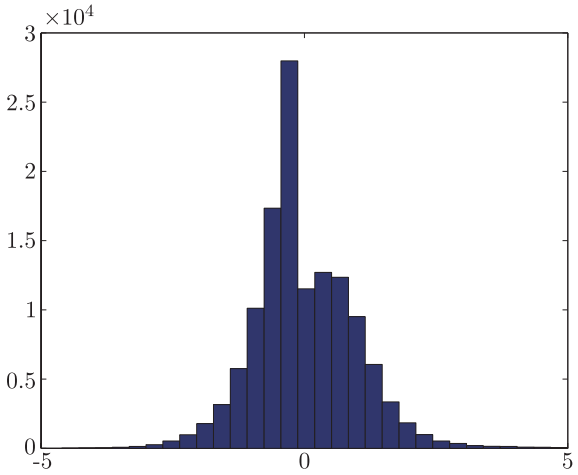

Figure 3. The histogram of the centered adjusted expression data.

levels of FDR as the thresholds for SC are the largest. From Theorems 2, 5, 8, and the simulation studies, we believe that the actual FDR resulting from using the BA and SC methods are closer to the nominal level than the NA method, and thus the SC method would be most reliable among all three procedures.

5. Discussion

We have investigated the effect of the non-normality on the cutting threshold and FDR control/estimation in multiple testing. Although we conservatively set $\pi_0 = 1$ in our asymptotic theory, our results can be generalized to incorporate other conservative estimates of π_0, such as that proposed by Storey (2002). The effect of inaccurate p-values on the estimation of π_0 is itself an interesting topic which is left as a future research project. As demonstrated in Theorems 1, 4, and 7, for the finite sample performance of the approximate FDR control methods, α is usually restricted to be away from 0 with a lower bound related to $n^{1/4}$. What happens if α is very close to 0 as in the sparse signal detection? Is there a more accurate approximation method that can relax the restriction on α? We expect that the saddlepoint approximation to the t-statistics would be a good alternative, and leave this problem to our future research.

We assumed independence structure in the data, but believe that our results can be extended to the situation where some weak dependence structure is imposed, such as the conditional independence structure assumed in Wu (2008). The fundamental fluctuation equation (A.1) and Lemma 1 are not much affected under weak dependence. For the general dependent case, a possible way to deal with the correlation among individual test statistics is via factor modeling so that the individual t-test statistics can be constructed based on the factor-corrected

data that are approximately independent, or at most weakly dependent, since the factor-corrected data is nearly only related to the idiosyncratic or firm-specific error in econometrics, see Fan, Han, and Gu (2012) and the references therein.

Acknowledgement

The authors are grateful to the Co-Editor, an associate editor, and two anonymous referees for comments and suggestions that led to substantial improvements in the paper. Jing's research was partially supported by Hong Kong RGC HKUST6019/10P, HKUST6019/12P, and HKUST6022/13P. Kong's research was supported in part by National NSFC No.11201080 and Humanity and Social Science Youth Foundation of Chinese Ministry of Education No. 12YJC910003. Zhou's research was supported in part by grant R-155-000-139-112 at the National University of Singapore.

Appendix: Proofs

Lemma A.1. Under the conditions in Theorem 1, $t_\alpha^N(m) = t_\alpha^N + O_p(m^{-1/2})$; Under the conditions in Theorem 4, $t_\alpha^b(m) = t_\alpha^b + O_p(m^{-1/2})$; Under the conditions in Theorem 7, $t_\alpha^c(m) = t_\alpha^c + O_p(m^{-1/2})$. If $w(t_\alpha) < \alpha^{-1} < w(0)$, then $t_\alpha(m) = t_\alpha + O_p(1/\sqrt{m})$, where $w(t)$ is the p.d.f. of P_1.

Proof. We only prove the first of three since the others can be proved similarly. Let $W^N(t)$ and $W_m^N(t)$ be the distribution and empirical distribution of the P_i^N's, respectively. By Theorem 2.11 of Stute (1982),

$$\sup_{|u|\le c/\sqrt{m}} \left\| W_m^N(t+u) - W_m^N(t) - (W^N(t+u) - W^N(t)) \right\| = O(m^{-3/4}\log(m)) \tag{A.1}$$

almost surely. Now we prove $Pr(t_\alpha^N(m) < t_\alpha^N + C/\sqrt{m}) \to 1$ for some constant C. By the definition of $t_\alpha^N(m)$ and (A.1), it is equivalent to show that

$$Pr\left(\frac{t_\alpha^N + C/\sqrt{m}}{W_m^N(t_\alpha^N + C/\sqrt{m})} > \alpha \right) \to 1, \tag{A.2}$$

for large enough m. By (A.1) and a Taylor expansion, it suffices to show that

$$Pr\left(t_\alpha^N + \frac{C}{\sqrt{m}} > \alpha W_m^N(t_\alpha^N) + \alpha w^N(t_\alpha^N)\frac{C}{\sqrt{m}} + O\Big(\big(\frac{C}{\sqrt{m}}\big)^2 \Big) \right) \to 1. \tag{A.3}$$

Since $W_m^N(t_\alpha^N) = W^N(t_\alpha^N) + O_P(1/\sqrt{m})$ and $\alpha W^N(t_\alpha^N) = t_\alpha^N$, (A.3) holds if $w^N(t_\alpha^N) < 1/\alpha$. Similarly, we can prove that $Pr(t_\alpha^N(m) > t_\alpha^N - C/\sqrt{m}) \to 1$.

Proof of Theorem 1. By Lemma A.1, we have $t_\alpha(m) - t_\alpha = O_p(1/\sqrt{m})$ and $t_\alpha^N(m) - t_\alpha^N = O_p(1/\sqrt{m})$. Then it suffices to show that

$$t_\alpha - t_\alpha^N = C_1(\alpha, t_\alpha^N, G_0^{-1}(1-t_\alpha^N))(\Phi^{-1}(1-t_\alpha^N) - G_0^{-1}(1-t_\alpha^N))(1+o_n(1)). \tag{A.4}$$

By (A.5) of Delaigle, Hall, and Jin (2010), we have

$$G_0^{-1}(1-t_\alpha^N)-\Phi^{-1}(1-t_\alpha^N)$$
$$=-\frac{1}{3}\gamma n^{-1/2}(\Phi^{-1}(1-t_\alpha^N))^2+\theta(n,\alpha)\{n^{-1/2}+(1+\Phi^{-1}(1-t_\alpha^N))^3 n^{-1}\}, \quad \text{(A.5)}$$

where $\theta(n,\alpha)$ satisfies $|\theta(n,\alpha)|\le C(B)$ (C(B) is a bounded positive constant) uniformly for all α such that $0\le\Phi^{-1}(1-t_\alpha^N)\le Bn^{1/4}$.

We next show that $H(t_\alpha^N)$ and $H(t_\alpha)$ are close enough, which implies that t_α^N and t_α are close enough. By the property of $f(u)$ and the Mean Value Theorem, we have

$$\begin{aligned}H(t_\alpha^N)-H(t_\alpha)&=H(t_\alpha^N)-\alpha=H(t_\alpha^N)-H^N(t_\alpha^N)\\&=\frac{t_\alpha^N(\overline{F}(\Phi^{-1}(1-t_\alpha^N))-\overline{F}(G_0^{-1}(1-t_\alpha^N)))}{\overline{F}(G_0^{-1}(1-t_\alpha^N))\overline{F}(\Phi^{-1}(1-t_\alpha^N))}\\&=\frac{\alpha}{\overline{F}(G_0^{-1}(1-t_\alpha^N))}(\overline{F}(\Phi^{-1}(1-t_\alpha^N))-\overline{F}(G_0^{-1}(1-t_\alpha^N)))\\&=\frac{\alpha(1+o_n(1))}{\overline{F}(G_0^{-1}(1-t_\alpha^N))}f(\Phi^{-1}(1-t_\alpha^N))\left(G_0^{-1}(1-t_\alpha^N)-\Phi^{-1}(1-t_\alpha^N)\right). \quad \text{(A.6)}\end{aligned}$$

(A.5) and (A.6) imply that $t_\alpha^N-t_\alpha=o_n(1)$. Letting $\xi\in(t_\alpha\wedge t_\alpha^N, t_\alpha^N\vee t_\alpha)$, we have by the Mean Value Theorem and the condition on $H'(t)$,

$$\begin{aligned}t_\alpha-t_\alpha^N&=\frac{H(t_\alpha)-H(t_\alpha^N)}{H'(\xi)}=\frac{H(t_\alpha)-H(t_\alpha^N)}{H'(t_\alpha^N)}(1+o_n(1))\\&=\left(H(t_\alpha)-H(t_\alpha^N)\right)\frac{(1+o_n(1))\overline{F}^2(G_0^{-1}(1-t_\alpha^N))g_0(G_0^{-1}(1-t_\alpha^N))}{g_0(G_0^{-1}(1-t_\alpha^N))\overline{F}(G_0^{-1}(1-t_\alpha^N))-t_\alpha^N f(G_0^{-1}(1-t_\alpha^N))}. \quad \text{(A.7)}\end{aligned}$$

Combining (A.5), (A.6), and (A.7) completes the proof of (A.4).

Proof of Theorem 2. By the definition of $t_\alpha(m)$, we have

$$\begin{aligned}\widehat{FDR}(t_\alpha^N(m))&=\widehat{FDR}(t_\alpha^N(m))-\widehat{FDR}(t_\alpha(m))+\alpha\\&=\alpha-\Big(\widehat{FDR}(t_\alpha(m))-\widehat{FDR}(t_\alpha)+\widehat{FDR}(t_\alpha)-\widehat{FDR}(t_\alpha^N)\\&\quad+\widehat{FDR}(t_\alpha^N)-\widehat{FDR}(t_\alpha^N(m))\Big). \quad \text{(A.8)}\end{aligned}$$

We investigate the terms in (A.8) one by one. By (A.1),

$$\begin{aligned}\widehat{FDR}(t_\alpha(m))-\widehat{FDR}(t_\alpha)&=\alpha-\frac{mt_\alpha}{R(t_\alpha)}=\frac{\alpha\overline{F}(G_0^{-1}(1-t_\alpha))-t_\alpha}{\overline{F}(G_0^{-1}(1-t_\alpha))}+O_P(\frac{1}{\sqrt{m}})\\&=O_P(\frac{1}{\sqrt{m}}). \quad \text{(A.9)}\end{aligned}$$

By (A.1) and Lemma 2,

$$\begin{aligned}
&\widehat{FDR}(t_\alpha^N) - \widehat{FDR}(t_\alpha^N(m)) \\
&\quad = \frac{mt_\alpha^N}{R(t_\alpha^N)} - \frac{mt_\alpha^N(m)}{R(t_\alpha^N(m))} \\
&\quad = \frac{t_\alpha^N}{\overline{F}(g_{\alpha,N})} - \frac{t_\alpha^N(m)}{\overline{F}(g_{\alpha,N,m})} + O_P(\frac{1}{\sqrt{m}}) \\
&\quad = \frac{(t_\alpha^N - t_\alpha^N(m))\overline{F}(g_{\alpha,N,m}) + t_\alpha^N(m)(\overline{F}(g_{\alpha,N,m}) - \overline{F}(g_{\alpha,N}))}{\overline{F}(g_{\alpha,N})\overline{F}(g_{\alpha,N,m})} \\
&\quad = O_p(\frac{1}{\sqrt{m}}),
\end{aligned} \tag{A.10}$$

where $g_{\alpha,N,m} = G_0^{-1}(1 - t_\alpha^N(m)), g_{\alpha,N} = G_0^{-1}(1 - t_\alpha^N)$. For the middle term in (A.8), we have

$$\begin{aligned}
&\widehat{FDR}(t_\alpha) - \widehat{FDR}(t_\alpha^N) \\
&\quad = \frac{mt_\alpha}{R(t_\alpha)} - \frac{mt_\alpha^N}{R(t_\alpha^N)} + O_p(\frac{1}{\sqrt{m}}) \\
&\quad = \frac{(t_\alpha - t_\alpha^N)\overline{F}(g_{\alpha,N}) + t_\alpha^N(\overline{F}(g_{\alpha,N}) - \overline{F}(G_0^{-1}(1 - t_\alpha)))}{\overline{F}(G_0^{-1}(1 - t_\alpha))\overline{F}(g_{\alpha,N})} + O_p(\frac{1}{\sqrt{m}}) \\
&\quad = \frac{(t_\alpha - t_\alpha^N)(g_0(g_{\alpha,N})\overline{F}(g_{\alpha,N}) - t_\alpha^N f(g_{\alpha,N}))(1 + o_n(1))}{g_0(g_{\alpha,N})\overline{F}(g_{\alpha,N})\overline{F}(G_0^{-1}(1 - t_\alpha))} + O_p(\frac{1}{\sqrt{m}}).
\end{aligned} \tag{A.11}$$

By (A.6) and (A.7),

$$\begin{aligned}
&\widehat{FDR}(t_\alpha) - \widehat{FDR}(t_\alpha^N) \\
&\quad = \frac{\alpha f(\Phi^{-1}(1 - t_\alpha^N))}{\overline{F}(G_0^{-1}(1 - t_\alpha))}(-G_0^{-1}(1 - t_\alpha^N) + \Phi^{-1}(1 - t_\alpha^N))(1 + o_n(1)).
\end{aligned} \tag{A.12}$$

By (A.8), (A.9), (A.12), (A.10), and (A.5) we have proved the theorem.

Proof of Theorem 3. By (A.1), Lemma A.1, and (A.1) of Delaigle, Hall, and Jin (2010),

$$\begin{aligned}
&FDP_m^N(t_\alpha^N(m)) \\
&= \frac{\sum_{i=1}^m I(H_i = 0,\ P_i^N < t_\alpha^N(m))}{\sum_{i=1}^m I(P_i^N < t_\alpha^N(m))} = \frac{\sum_{i=1}^m I(H_i = 0,\ P_i^N < t_\alpha^N)}{\sum_{i=1}^m I(P_i^N < t_\alpha^N)} + O_P(\frac{1}{\sqrt{m}}) \\
&= \frac{\pi_0 Pr\left(\tilde{T}_i > \Phi^{-1}(1 - t_\alpha^N)\right)}{Pr\left(P_1^N < t_\alpha^N\right)} + O_P(\frac{1}{\sqrt{m}}) \\
&= \pi_0 \alpha \exp\left(-1/3\gamma n^{-1/2}[\Phi^{-1}(1 - t_\alpha^N)]^3\right)
\end{aligned}$$

$$\times \Big(1+\theta(n,\alpha)[(1+\Phi^{-1}(1-t_\alpha^N))n^{-1/2}+(1+\Phi^{-1}(1-t_\alpha^N))^4 n^{-1}]\Big)$$
$$+O_P(1/\sqrt{m}). \tag{A.13}$$

Proof of Theorem 4. Let $\hat{\gamma}_n(i)$ and $\tilde{\gamma}_n(i)$ be the sample skewness and kurtosis based on the observations for gene i with the divisor equal to n, respectively. By (A.8) of Delaigle, Hall, and Jin (2010), we have

$$\begin{aligned} &G_0^{-1}(1-t_\alpha^b)-(G_b^i)^{-1}(1-t_\alpha^b)\\ &\quad=-\frac{1}{3}(\Phi^{-1}(1-t_\alpha^b))^2(\gamma-\hat{\gamma}_n(i))n^{-1/2}\\ &\qquad-\Theta(n,\alpha)\{n^{-1/2}+(1+\Phi^{-1}(1-t_\alpha^b))^3 n^{-1}\}, \end{aligned} \tag{A.14}$$

where $\Theta(n,\alpha)$ satisfies $|\Theta(n,\alpha)|\le C(B)$ for all $\mathcal{X}_i$ with $S_i>1/2$ and $\tilde{\gamma}_n(i)\le B$, uniformly for the distributions of X satisfying $EX=0$, $EX^2=1$, and $EX^4\le B$, and for α such that $0\le\Phi^{-1}(1-t_\alpha^b)\le Bn^{1/4}$.

We investigate how close $H(t_\alpha)$ is to $H(t_\alpha^b)$. Recall that $G_b:=G_b^1$. By some calculations we have

$$H(t_\alpha)-H(t_\alpha^b)=\frac{\alpha}{\overline{F}(G_0^{-1}(1-t_\alpha^b))}(\overline{F}(G_0^{-1}(1-t_\alpha^b))-E\overline{F}(G_b^{-1}(1-t_\alpha^b)))), \tag{A.15}$$

with

$$\begin{aligned} &\overline{F}(G_0^{-1}(1-t_\alpha^b))-E\overline{F}(G_b^{-1}(1-t_\alpha^b)))\\ &\quad=E[\overline{F}(G_0^{-1}(1-t_\alpha^b))-\overline{F}(G_b^{-1}(1-t_\alpha^b))](I(S_1>\frac{1}{2})+I(S_1\le\frac{1}{2})). \end{aligned} \tag{A.16}$$

By Markov's inequality,

$$Pr(S_1\le\frac{1}{2})\le Pr(|S_1^2-1|\ge\frac{3}{4})\le\frac{C(B)}{n}. \tag{A.17}$$

By the Dominance Convergence Theorem, we have

$$Pr(\tilde{\gamma}_n(1)\le B)-1=o_n(1). \tag{A.18}$$

From (A.16), (A.17), and (A.18),

$$\begin{aligned} &\overline{F}(G_0^{-1}(1-t_\alpha^b)-E\overline{F}(G_b^{-1}(1-t_\alpha^b)))\\ &\quad=E[\overline{F}(G_0^{-1}(1-t_\alpha^b)-\overline{F}(G_b^{-1}(1-t_\alpha^b)))]I(S_1>\frac{1}{2})I(\tilde{\gamma}_n(1)\le B)(1+o_n(1)). \end{aligned} \tag{A.19}$$

Let ζ be some random variable taking values in $(G_0^{-1}(1-t_\alpha^b)\wedge G_b^{-1}(1-t_\alpha^b), G_0^{-1}(1-t_\alpha^b)\vee G_b^{-1}(1-t_\alpha^b))$. By (A.14),

$$\overline{F}(G_0^{-1}(1-t_\alpha^b)-E\overline{F}(G_b^{-1}(1-t_\alpha^b)))$$

$$= Ef(\zeta)\left(G_0^{-1}(1-t_\alpha^b) - G_b^{-1}(1-t_\alpha^b)\right) I(S_1 > \frac{1}{2}) I(\tilde{\gamma}_n(1) \le B)(1+o_n(1))$$
$$= Ef(G_0^{-1}(1-t_\alpha^b))\left(G_0^{-1}(1-t_\alpha^b) - G_b^{-1}(1-t_\alpha^b)\right)$$
$$\times I(S_1 > 1/2) I(\tilde{\gamma}_n(1) \le B)(1+o_n(1)). \quad \text{(A.20)}$$

On the other hand,

$$t_\alpha^b - t_\alpha = \frac{H(t_\alpha^b) - H(t_\alpha)}{H'(t_\alpha^b)}(1+o_n(1))$$
$$= \frac{(H(t_\alpha^b) - H(t_\alpha))\overline{F}^2(G_0^{-1}(1-t_\alpha^b)) g_0(G_0^{-1}(1-t_\alpha^b))}{\overline{F}(G_0^{-1}(1-t_\alpha^b)) g_0(G_0^{-1}(1-t_\alpha^b)) - t_\alpha^b f(G_0^{-1}(1-t_\alpha^b))}(1+o_n(1)). \quad \text{(A.21)}$$

Substituting (A.20) and (A.14) into (A.15), and then substituting (A.15) into (A.21) proves the theorem.

Proof of Theorem 5. The proof of Theorem 5 is similar to that of Theorem 2.

Proof of Theorem 6. The proof of Theorem 6 is similar to that of Theorem 3, if we use (A.2) instead of (A.1) of Delaigle, Hall, and Jin (2010).

Proof of Theorem 7. (2.14) implies that

$$\Phi_c^{-1}(v) = \Phi^{-1}(v)\left(1 - \frac{1}{3}(\hat{\gamma}_m) n^{-1/2}\Phi^{-1}(v) \right.$$
$$\left. + \tilde{\theta}(n,\alpha)\{(1+\Phi^{-1}(v))^{-1} n^{-1/2} + (1+\Phi^{-1}(v))^2 n^{-1}\}\right), \quad \text{(A.22)}$$

where $\tilde{\theta}(n,\alpha) \le C(B)$, a bounded constant for v satisfying $0 \le 1 - \Phi^{-1}(v) \le Bn^{1/4}$. The proof of the present theorem is the same as that of Theorem 1, except for replacing t_α^N there by t_α^c, and $G_0^{-1}(\cdot) - \Phi^{-1}(\cdot)$ by $G_0^{-1}(\cdot) - \Phi_c^{-1}(\cdot)$.

Proof of Theorem 8. The proof of Theorem 8 is similar to that of Theorem 2.

Proof of Theorem 9. The proof of Theorem 9 is similar to that of Theorem 3.

References

Abbas, A., Kong, X. B., Liu, Z., Jing, B. Y. and Gao, X. (2013). Automatic peak selection by a Benjamini-Honchberg-based algorithm. *PLoS ONE* **7**, DOI: 10.1371/journal.pone.0053112.

Bajgrowicz, P. and Scaillet, O. (2012). Technical trading revisited: false discoveries, persistence tests, and transaction costs. *J. Finan. Econ.* **106**, 473-491.

Barras, L., Acaillet, O. and Wermers, R. (2010). False discoveries in mutual fund performance: measuring luck in estimated alphas. *J. Finan.* **LXV-1**, 179-216.

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.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.* **25**, 60-83.

Benjamini, Y., Krieger, M. A. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491-507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165-1188.

Briec, W., Kerstens, K. and Jokung, O. (2007). Mean-variance-skewness portfolio performance gauging: a general shortage function and dual approach. *Management Science* **53**, 135-149.

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. University of Melbourne, Melbourne. (Available from `http://arxiv.org/`.)

Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on student's t-statistics. *J. Roy. Statist. Soc. Ser B.* **73**, 283-301.

Fan, J. Q., Hall, P. and Yao, Q. (2007). To how many simultaneous hypothesis tests can normal, student's t or bootstrap calibration be applied? *J. Amer. Stat. Assoc.* **102**, 1282-1288.

Fan, J., Han, X. and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107**, 1019-1035.

Hu, J., Zhao, H. and Zhou, H. (2011). False discovery rate control with groups. *J. Amer. Statist. Assoc.* **105**, 1215-1227.

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

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *J. Roy. Statist. Soc. Ser B.* **74**, 163-182.

Liu, W. and Shao, Q. (2013). A Cramér moderate deviation theorem for Hotelling's T^2 statistic with applications to global tests. *Ann. Statist.* **41**, 296-322.

Nettleton, D., Hwang, J., Caldo, R. and Wise, R. (2006). Estimating the number of true null hypotheses from a histogram of p values. *J. Agric. Biol. Environ. Statist.* **11**, 337-356.

Sentana, E. (2008). The econometrics of mean-variance efficiency tests: a survey. *The Econometrics J.* **12**, C65-C101.

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

Stute, W. (1982). The oscillation behavior of empirical processes. *Ann. Prob.* **10**, 86-107.

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.

Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

E-mail: majing@ust.hk

Department of Mathematical Sciences, Soochow University, Soochow, Jiangsu Province 215021, China.

E-mail: kongxblqh@gmail.com

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546.

E-mail: stazw@nus.edu.sg

(Received February 2013; accepted January 2014)