

To How Many Simultaneous Hypothesis Tests Can Normal, Student's t or Bootstrap Calibration Be Applied?

Jianqing FAN, Peter HALL, and Qiwei YAO

In the analysis of microarray data, and in some other contemporary statistical problems, it is not uncommon to apply hypothesis tests in a highly simultaneous way. The number, N say, of tests used can be much larger than the sample sizes, n , to which the tests are applied, yet we wish to calibrate the tests so that the overall level of the simultaneous test is accurate. Often the sampling distribution is quite different for each test, so there may not be an opportunity to combine data across samples. In this setting, how large can N be, as a function of n , before level accuracy becomes poor? Here we answer this question in cases where the statistic under test is of Student's t type. We show that if either the normal or Student t distribution is used for calibration, then the level of the simultaneous test is accurate provided that $\log N$ increases at a strictly slower rate than $n^{1/3}$ as n diverges. On the other hand, if bootstrap methods are used for calibration, then we may choose $\log N$ almost as large as $n^{1/2}$ and still achieve asymptotic-level accuracy. The implications of these results are explored both theoretically and numerically.

KEY WORDS: Bonferroni's inequality; Edgeworth expansion; Genetic data; Large-deviation expansion; Level accuracy; Microarray data; Quantile estimation; Skewness; Student's t statistic.

1. INTRODUCTION

Modern technology allows us to collect a large amount of data in one scan of images. This is exemplified in genomic studies using microarrays, tiling arrays, and proteomic techniques. In the analysis of microarray data, and in some other contemporary statistical problems, we often wish to make statistical inference simultaneously for **all important parameters**. The number of parameters, N , is frequently much larger than the sample size, n . Indeed, the sample size is typically small; for example, $n = 8, 20$, and 50 are considered typical, moderately large, and large for microarray data. In other situations such as the analysis of single nucleotide polymorphisms (SNPs), the sample size n can be several hundred, whereas the number of SNPs, N , can be in the order of 10^5 . The question arises naturally as to how large N can be before the accuracy of simultaneous statistical inference becomes poor.

Important results in this direction have been obtained by van der Laan and Bryan (2001), who showed that the population mean and variance parameters can be consistently estimated when $\log N = o(n)$ if the observed data are bounded. Bickel and Levina (2004) gave similar results in a high-dimensional classification problem; Fan, Peng, and Huang (2005b) and Huang, Wang, and Zhang (2005) studied semiparametric inference where $N \rightarrow \infty$; and Hu and He (2007) proposed an enhanced quantile normalization based on high-dimensional singular value decomposition to reduce information loss in gene expression profiles.

Korosok and Ma (2007) treated the problem of uniform, simultaneous estimation of a large number of marginal distributions, showing that if $\log N = o(n)$, and if certain other conditions hold, then $\max_{1 \leq i \leq N} \|\hat{F}_i - F_i\|_\infty \rightarrow 0$, where \hat{F}_i is an estimator of the i th marginal distribution F_i . As a corollary, they proved that a p value, \hat{P}_i , of \hat{F}_i converges uniformly in i to its counterpart P_i for F_i , provided that $\log N = o(n^{1/2})$ namely,

$$\max_{1 \leq i \leq N} \|\hat{P}_i - P_i\|_\infty \rightarrow 0. \quad (1)$$

These results are important advances in the literature of simultaneous testing, where p values are popularly assumed to be known. Examples in the latter setting have been given by Benjamini and Yekutieli (2001), Sarkar (2002, 2006), Dudoi, Shaffer, and Boldrick (2003), Donoho and Jin (2004), Efron (2004), Genovese and Wasserman (2004), Storey, Taylor, and Siegmund (2004), Lehmann and Romano (2005), and Lehmann, Romano, and Shaffer (2005), where many new ideas have been introduced to control different aspects of simultaneous hypothesis testing and false discovery rate (FDR).

In many practical settings, the assumption that p values are calculated without error is unrealistic, but it is unclear how good the approximation must be for simultaneous inference to be feasible. **Simple consistency, as evidenced by (1), is not enough**; the level of accuracy required must increase with N . More precisely, letting α_N be the significant level, which tends to 0 as $N \rightarrow \infty$, the required accuracy is then

$$\max_{1 \leq i \leq N} \|\hat{P}_i - P_i\|_\infty = o(\alpha_N). \quad (2)$$

In this article we provide a concise solution to this problem. For example, we show that in the case of simultaneous t tests, calibrated by reference to the normal or Student t distribution, a necessary and sufficient condition for overall level accuracy to be asymptotically correct is (a) **$\log N = o(n^{1/3})$** . This is true even if the sampling distribution is highly skewed or heavy-tailed. On the other hand, if bootstrap methods are used for

Jianqing Fan is Frederick L. Moore'18 Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, and Director, Center for Statistical Research, Academy of Mathematics and Systems Science, Beijing, China (E-mail: jqfan@princeton.edu). Peter Hall is Professor, Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia (E-mail: halpstat@ms.unimelb.edu.au). Qiwei Yao is Professor, Department of Statistics, London School of Economics, London WC2A 2AE, U.K., and Guanghua School of Management, Peking University, China (E-mail: q.yao@lse.ac.uk). Fan's work was sponsored in part by National Science Foundation grants DMS-0354223 and DMS-0704337, National Institutes of Health grant R01-GM07261, and NSF grant 10628104 of China. Yao's work was sponsored in part by EPSRC grant EP/C549058. The authors thank the joint editors, associate editor, and two anonymous referees for their helpful comments that lead to the improvement of the manuscript.

estimating p values, then the asymptotic level of the simultaneous test is accurate **as long as (b) $\log N = o(n^{1/2})$** . These results make clear the advantages offered by bootstrap calibration. We explore them both numerically and theoretically. Result (a) needs only bounded third moments of the sampling distribution, although our proof of (b) requires more restrictions.

Take the case of the familywise error rate (FWER) as an example. If the overall error rate is controlled at p , then k_n hypotheses with the smallest p values are rejected, where

$$k_n = \max\{i : P_{(i)} \leq p/N, i = 1, \dots, N\}, \quad (3)$$

P_i denotes the significance level of the i th test, and $\{P_{(i)}\}$ are the ordered values of $\{P_i\}$. The approach of Benjamini and Hochberg (1995) to controlling false discovery rate (FDR) at p is to select $S_0 = \{i : P_i \leq P_{(k_n)}\}$, where

$$k_n = \max\{i : P_{(i)} \leq ip/N, i = 1, \dots, N\}. \quad (4)$$

If the distributions from which the P_i 's are computed need to be estimated, then, in view of (3) or (4), the error of the estimators \hat{P}_i should equal $o(N^{-1})$ to correctly sort $\{P_i\}$, and the approximation (1) requires significant refinement. Indeed, it corresponds to (2) with $\alpha_N = p/N$. This is a very stringent requirement ($\alpha_N = 10^{-5}$, if $p = .1$ and $N = 10^4$), and accuracy is difficult to achieve for many practical situations.

Sometimes standard statistical arguments provide attractive ways of selecting significant genes. For example, in their analysis of gene expression data, Fan, Tam, Vande Woude, and Ren (2004) took $\alpha_N = .001$ and found the significant set of genes,

$$S = \{i : P_i \leq \alpha_N, i = 1, \dots, N\}, \quad (5)$$

for $N = 15,000$ simultaneous tests. Here α_N is an order of magnitude larger than N^{-1} , and the approximation errors when estimating p values need be $o(\alpha_N)$ only when computing (5), rather than $o(N^{-1})$ in the FWER problem. The approximate equivalence between the set S in (5) and S_0 with $\hat{p} = N\alpha_N/\#(S)$ has been given by Fan, Chen, Chan, Tam, and Ren (2005a) and an earlier draft of this article. Thus our results have implications for controlling FDRs when p values are approximated.

Results such as these, addressing the case of independent tests, also have implications under the strong assumption of positive regression dependency (see Benjamini and Yekutieli 2001, p. 1170, for a discussion). The results are also pertinent to many other dependent tests regardless of the sign of dependence; for example, Clarke and Hall (2007) have shown that if test statistics can be modeled by linear and related processes, and also in cases where the statistics are computed from linear-process data (including tests based on the Student t statistic calculated from such data), then both FDR and FWER are asymptotically identical to their counterparts in the case of independence.

The key requirement here is that the distributions of individual test statistics have very light tails; the tails of t statistics, for moderately large group sizes, are adequate. In such instances, high-level exceedences occur only because the independent disturbances that produce the dependent process are propitiously aligned. In light-tailed cases, it is very unlikely that a particular disturbance will be so large as to carry several nearby values of the test statistic over the critical point, and when the index, i ,

of the test statistic is altered (to, e.g., an adjacent value), the alignment changes so much that a chance-level exceedence is very unlikely to be repeated. Cases with sufficiently light tails include those in which tests are based on Student t statistics with $\log N = o(n)$, and the underlying distributions have three finite moments. Hall and Wang (2007) showed that in this setting there exists a strong approximation in which the pattern of exceedences over a high level is identical, with probability converging to 1 as $N \rightarrow \infty$, to what would occur if the test statistics were truly independent.

Therefore, the stochastic process of false discoveries, which in general is a cluster-type process, is first-order equivalent to its counterpart in the case of independence, that is, to a homogeneous Poisson process. In other, heavy-tailed cases, modeling the extent of dependence is attractive (see, e.g., Efron 2007), but there too one would want to start with accurate marginal p values. We address that topic here.

Practical implications of our work include the following:

- a. When calibrating multiple hypothesis tests using the Student t statistic, for example with a view to controlling FWER, impressive level accuracy can be obtained through a Student t approximation. Only a mild moment condition is needed.
- b. Nevertheless, accuracy is noticeably improved by using the bootstrap.
- c. These results also apply in the case of generalized FWER.
- d. Owing to the limitation on the number of simultaneous hypotheses that can be accurately tested for a given n and α_N , other methods of prescreening are needed when there are excessively many hypotheses to be tested.

The methodologies that we consider are confined to the Student t statistic, because recent, remarkably general results that require only low-order moment assumptions are available in that case. Indeed, this distinguishes the Student t context from simpler but less practically relevant settings in which even necessary conditions for results are very onerous (see Shao 1999; Wang 2005 for discussion). However, although it will take time for theoretical research to catch up to the advance of statistical methodologies for analyzing microarray data, it seems very likely that analogs of the detailed properties given in this article also hold for other resampling methods based on pivoted, or self-normalized, statistics.

In general, the resampling of microarray data can proceed in any of several different ways. It can use permuted, pooled data; it can involve resampling within each group and basing inference on the bootstrap distribution of the resulting pivoted statistic; or it can use a combination of both data pooling and residual resampling, after testing for the equality of between-group distributions. We study the second of these three approaches in this work; although the first and third, under mild regularity conditions, require theory to which we do not have access at present, it seems likely that conclusions b and c given two paragraphs earlier apply there as well. Extended versions of the second approach, in which the data within a group were something other than independent and identically distributed, would require a model-based approach but again generally would be amenable to treatment using the arguments that we give here.

In a sequence of articles, Finner and Roters (1998, 1999, 2000, 2001, 2002) developed theoretical properties of n simultaneous hypothesis tests as n increases. But their work differs from ours in a major respect, through their assumption that the exact significance level can be tuned to a known value in the continuum. In such cases, there is no theoretical limit to how large nN can be. In comparison, in the setting of the genetic problems that motivates our work, level inaccuracies limit the effective size of N . We delineate this limitation using both theoretical and numerical arguments.

The article is organized as follows. In Section 2 we formulate the accuracy problem for simultaneous tests and also outline statistical models and testing procedures. We present our main results in Section 3, where we answer the question of how many hypotheses can be tested simultaneously without the overall significance level being seriously in error. The theoretical definition of the latter property is that the overall significance level should converge to the nominal one as the number of tests increases if the samples on which the tests are based are independent, and that the limiting level should not exceed the nominal one if the independence condition is violated and a Bonferroni bound is used.

We present numerical investigations among various calibration methods in Section 4. Technical proofs of the results given in Section 3 are relegated to Section 5.

2. MODEL AND METHODS FOR TESTING

2.1 Basic Model and Methodology

The simplest model is that in which we observe random variables

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad 1 \leq i < \infty, 1 \leq j \leq n, \quad (6)$$

with the index i denoting the i th gene, j indicating the j th array, and the constant μ_i representing the mean effect for the i th gene. We assume the following:

For each i , $\epsilon_{i1}, \dots, \epsilon_{in}$ are independent and identically distributed random variables with 0 expected value. (7)

The results given herein are readily extended to the case where $n = n_i$ depends on i , but taking n fixed simplifies our discussion.

Let $T_i = n^{1/2} \bar{Y}_i / S_i$, where

$$\bar{Y}_i = \frac{1}{n} \sum_{j=1}^n Y_{ij} \quad \text{and} \quad S_i^2 = \frac{1}{n} \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2.$$

For a given value of i , we wish to test the null hypothesis H_{0i} that $\mu_i = 0$ against the alternative hypothesis H_{1i} that $\mu_i \neq 0$, for $1 \leq i \leq N$, say. We first study this standard testing problem of **controlling the probability of making at least one false discovery**, which requires calculating p values with accuracy $o(N^{-1})$, the same as that needed in (3). We then extend our results to control the relaxed FDR in (5), which is less stringent.

A standard test is to reject H_{0i} if $|T_i| > t_\alpha$. Here t_α denotes the solution of equation

$$P(|Z| > t_\alpha) = 1 - (1 - \alpha)^{1/N} \quad \text{or} \quad (8)$$

$$P\{|T(n-1)| > t_\alpha\} = 1 - (1 - \alpha)^{1/N},$$

where Z and $T(k)$ have the standard normal distribution and Student t distribution with k degrees of freedom. Note that (8) serves only to give a definition of t_α that is commonly used in practice; it does not amount to an assumption about the sampling distribution of the data. Indeed, $\alpha_N = 1 - (1 - \alpha)^{1/N}$ and α are a one-to-one map. The core of the argument in this article is that the accuracy of the distributional approximations implicit in (8) are based on a delicate relationship between n and N , which is central to the question of how many simultaneous tests are possible.

2.2 Significance Levels for Simultaneous Tests

If H_{0i} is true, then the significance level of the test restricted to gene i is given by

$$p_i = P_{0i}(|T_i| > t_\alpha), \quad (9)$$

where P_{0i} denotes probability calculated under H_{0i} . For the approach (3), which is nearly the same as the Benjamini–Hochberg method, we ask how fast N can diverge to infinity without upsetting the condition

$$\max_{1 \leq i \leq N} p_i = o(1) \quad \text{and} \quad \sum_{i=1}^N p_i = \beta + o(1), \quad (10)$$

for some $0 < \beta < \infty$. The importance of (10) is that it implies that the significance level of the simultaneous test, described in Section 2.1, is

$$\alpha(N) \equiv P(H_{0i} \text{ rejected for at least one } i \text{ in the range } 1 \leq i \leq N) \quad (11)$$

$$\leq \sum_{i=1}^N p_i = \beta + o(1). \quad (12)$$

Therefore (10) guarantees asymptotic conservatism of level as $N \rightarrow \infty$. If, in addition to (7), we assume that

$$\text{the sets of variables } \{\epsilon_{ij}, 1 \leq j \leq n\} \text{ are independent for different } i, \quad (13)$$

then (10) guarantees asymptotic exactness of level [provided that we take $\beta = -\log(1 - \alpha)$] as $N \rightarrow \infty$. Assuming (7) and (10),

$$\text{if (13) holds, then } \alpha(N) = 1 - e^{-\beta} + o(1), \quad (14)$$

where $\alpha(N)$ is as defined in (11). Result (14) also holds, with the identity $\alpha(N) = 1 - e^{-\beta} + o(1)$ replaced by $\alpha(N) \leq 1 - e^{-\beta} + o(1)$, if (13) is replaced by the positive regression dependency assumption (Benjamini and Yekutieli 2001, p. 1170).

2.3 Generalized Familywise Error Rate

The foregoing results can be generalized by extending the definition at (11) to

$$\alpha_k(N) \equiv P(H_{0i} \text{ rejected for at least } k \text{ values of } i \text{ in the range } 1 \leq i \leq N)$$

$$\leq \frac{1}{k} \sum_{i=1}^N p_i = k^{-1} \beta + o(1), \quad (15)$$

where (15) follows from (10). We also have, under (7) and (10), the following analog of (14):

$$\text{if (13) holds, then } \alpha_k(N) = 1 - \sum_{j=1}^k \frac{\beta^j}{j!} e^{-\beta} + o(1). \quad (16)$$

Under the positive regression dependency assumption, the equality in (16) would be replaced by \leq .

2.4 Methods for Calibration

For calibration against the normal or Student t distributions, we take the critical point t_α to be the solution of the respective equations (8). Here we consider bootstrap calibration; Edgeworth correction (see, e.g., Hall 1990) also could be used.

Let $Y_{i1}^\dagger, \dots, Y_{in}^\dagger$ denote a bootstrap resample drawn by sampling randomly, with replacement, from $\mathcal{Y}_i = \{Y_{i1}, \dots, Y_{in}\}$. Put $Y_{ij}^* = Y_{ij}^\dagger - \bar{Y}_i$ and $T_i^* = n^{1/2} \bar{Y}_i^* / S_i^*$, where $\bar{Y}_i^* = n^{-1} \times \sum_j Y_{ij}^*$ and $(S_i^*)^2 = n^{-1} \sum_j (Y_{ij}^* - \bar{Y}_i^*)^2$. Write z_α for the conventional normal critical point for N simultaneous tests; that is, z_α solves $P(|Z| > z_\alpha) = 1 - (1 - \alpha)^{1/N}$. (We also could use the Student t point.) Define $\xi = \hat{f}_i(\alpha)$ to be the solution of

$$P(|T_i^*| > z_\xi \mid \mathcal{Y}_i) = 1 - (1 - \alpha)^{1/N}.$$

Our bootstrap critical point is $\hat{t}_{i\alpha} = z_{\hat{f}_i(\alpha)}$; we reject H_{0i} if and only if $|T_i| > \hat{t}_{i\alpha}$. Here the definition of p_i at (9) should be replaced by

$$p_i = P_{0i}(|T_i| > \hat{t}_{i\alpha}). \quad (17)$$

With this new definition, (14), (15), and (16) continue to be consequences of (10).

3. THEORETICAL RESULTS

3.1 Asymptotic Results

Define κ_{i3} to be the third cumulant or, equivalently, the skewness of the distribution of $\epsilon'_i = \epsilon_{i1}/(E\epsilon_{i1}^2)^{1/2}$. The novelty of Theorem 1 is that it gives a particularly concise account of the relationship between the validity of (10) for some value of β and the rates of growth of n and N . Corollary 1 converts this relationship into a necessary and sufficient condition [specifically, $\log N = o(n^{1/3})$] for (10) to hold for the particular value $\beta = -\log(1 - \alpha)$. This result is of direct interest for multiple hypothesis testing. Theorem 2 shows that if we use bootstrap calibration rather than calibration through the normal approximation or Student t distribution, then the earlier condition can be weakened to $\log N = o(n^{1/2})$. The practical ramifications of these results are explored in Section 3.2. They give direct access to practically important information about how large N can be for a given value of n , before the accuracy of the simultaneous test is seriously degraded.

Theorem 1. Assume that

$$\max_{1 \leq i \leq N} E|\epsilon'_i|^3 = O(1) \quad (18)$$

as $N \rightarrow \infty$, and suppose also that $N = N(n) \rightarrow \infty$ in such a manner that $(\log N)/n^{1/3} \rightarrow \gamma$, where $0 \leq \gamma < \infty$. Define t_α

by either of the formulas at (8), and p_i by (9). Then (10) holds with

$$\beta = \beta(N) \equiv -\frac{\log(1 - \alpha)}{N} \sum_{i=1}^N \cosh\left(\frac{1}{3} \gamma^3 \kappa_{i3}\right), \quad (19)$$

where $\cosh(x) = (e^x + e^{-x})/2$.

Note that $\beta(N)$, defined at (19), is bounded by $|\log(1 - \alpha)| \times \cosh(\gamma^3 B)$ uniformly in N , where $B = \sup_i |\kappa_{i3}|$.

Corollary 1. Assume the conditions of Theorem 1. If $\gamma = 0$ [i.e., if $\log N = o(n^{1/3})$], then (10) holds with $\beta = -\log(1 - \alpha)$, and if $\gamma > 0$, then (10) holds with $\beta = -\log(1 - \alpha)$ if and only if $N^{-1} \sum_{i \leq N} |\kappa_{i3}| \rightarrow 0$, that is, if and only if the limit of the average absolute values of the skewnesses of the distributions of $\epsilon_{11}, \dots, \epsilon_{N1}$ equals 0.

Because Corollary 1 implies (10), it also entails (12) and (14)–(16).

Theorem 2. Strengthen (18) to the assumption that for a constant $C > 0$, $P(|\epsilon'_i| \leq C) = 1$, and suppose also that $N = N(n) \rightarrow \infty$ in such a manner that $\log N = o(n^{1/2})$. Define $\hat{t}_{i\alpha} = z_{\hat{f}_i(\alpha)}$, as in Section 2.4, and define p_i by (17). Then (10) holds with $\beta = -\log(1 - \alpha)$.

The key conditions connecting n and N — $\log N = o(n^{1/3})$ in Theorem 1 and $\log N = o(n^{1/2})$ in Theorem 2—permit N to be exponentially larger than n (in particular, for N to be of larger order than any polynomial in n) before asymptotic level inaccuracies occur. This striking result is a consequence of the fact that a normal approximation to the distribution of a mean actually improves, in relative terms, as one moves further out into the tails.

3.2 Applications to Controlling Error Rate

Define t_α and $\hat{t}_{i\alpha}$ by (8) and as in Section 2.4. In the proof of Theorem 1 it is shown that, with $\beta = -\log(1 - \alpha)$ and using conventional calibration, $P_{0i}(|T_i| > t_\alpha) = \beta N^{-1} + o(N^{-1})$, uniformly in i under the null hypotheses, provided $\log N = o(n^{1/3})$; and that when using bootstrap calibration, $P_{0i}(|T_i| > \hat{t}_{i\alpha}) = \beta N^{-1} + o(N^{-1})$, again uniformly in i , if $\log N = o(n^{1/2})$. These results substantially improve a uniform convergence property of Korosok and Ma (2007), at the expense of more restrictions on N .

When the p values in (3) need to be estimated, the estimation errors should be of order $o(N^{-1})$, where N diverges with n . On the other hand, when p values in (5) are estimated, the precision can be of order $o(\alpha_N)$, where for definiteness we take $\alpha_N = CN^{-a}$ with $C > 0$ and $a \in (0, 1]$. In the latter case, the results in Theorems 1 and 2 continue to apply; there is no relaxation, despite the potential simplicity of the problem.

To appreciate why this is so, note that the tail probability of the standard normal distribution satisfies $P(|Z| \geq x) \sim \exp(-x^2/2)/(\sqrt{2\pi}x)$. Suppose that the large deviation result holds up to the point $x = x_n$, which should be of order $o(n^{1/6})$ for the Student t calibration and $o(n^{1/4})$ for bootstrap calibration. Setting $P(|Z| \geq x)$ equal to α_N yields $\log(1/\alpha_N) = \frac{1}{2}x_n^2 + \log x_n + \dots$, that is,

$$a \log N = \frac{1}{2}x_n^2 + \log x_n + \log(\sqrt{2\pi}C) + \text{smaller order terms.}$$

Regardless of the values of $C > 0$ and $a \in (0, 1]$, this relation implies that the condition $x_n = o(n^{1/6})$ is equivalent to $\log N = o(n^{1/3})$ and $x_n = o(n^{1/4})$ entails $\log N = o(n^{1/2})$, although taking a close to 0 will numerically improve approximations in both theory and practice.

4. NUMERICAL PROPERTIES

4.1 Simulation Study

Here we construct models that reflect aspects of gene expression data. Toward this end, we divide genes into three groups. Within each group, the genes share one unobserved common factor with different factor loadings. In addition, there is an unobserved common factor among all of the genes across the three groups. For simplicity of presentation, we assume that N is a multiple of three. We denote by Z_{ij} a sequence of independent $N(0, 1)$ random variables and by χ_{ij} a sequence of independent random variables of the same distribution as that of $(\chi_m^2 - m)/\sqrt{2m}$. Note that χ_{ij} has mean 0, variance 1, and skewness $\sqrt{8/m}$. In our simulation study, we set $m = 6$.

With given factor loading coefficients a_i and b_i , the error ϵ_{ij} in (6) is defined as

$$\epsilon_{ij} = \frac{Z_{ij} + a_{i1}\chi_{j1} + a_{i2}\chi_{j2} + a_{i3}\chi_{j3} + b_i\chi_{j4}}{(1 + a_{i1}^2 + a_{i2}^2 + a_{i3}^2 + b_i^2)^{1/2}},$$
$$i = 1, \dots, N, j = 1, \dots, n,$$

where $a_{ij} = 0$ except that $a_{i1} = a_i$ for $i = 1, \dots, \frac{1}{3}N$, $a_{i2} = a_i$ for $i = \frac{1}{3}N + 1, \dots, \frac{2}{3}N$, and $a_{i3} = a_i$ with $i = \frac{2}{3}N + 1, \dots, N$. Note that $E\epsilon_{ij} = 0$ and $\text{var}(\epsilon_{ij}) = 1$, and that the within-group correlation is in general stronger than the between-group correlation, because the former shares one extra common factor. We consider two specific choices of factor loadings: case I, where the factor loadings are taken to be $a_j = .25$ and $b_j = .1$ for all j (thus the ϵ_{ij} 's have the same marginal distribution, although they are correlated), and case II, for which the factor loadings a_i and b_i are generated independently from $U(0, .4)$ and $U(0, .2)$.

The “true” gene expression levels μ_i are taken from a realization of the mixture of a point mass at 0 and a double-exponential distribution: $c\delta_0 + \frac{1}{2}(1 - c)\exp(-|x|)$, where $c \in (0, 1)$ is a constant. With the noise and the expression level given earlier, Y_{ij} generated from (6) represents, for each fixed j , the observed log-ratios between the two-channel outputs of a c-DNA microarray. Note that $|\mu_j| \geq \log 2$ means that the true expression ratio exceeds 2. The probability, or the empirical fraction, of this event equals $\frac{1}{2}(1 - c)$.

For each α_N , we compute the p value according to the normal approximation, t -approximation, and bootstrap method. Because the marginal null distributions of T_i are the same for $i = 1, \dots, N$, we also can average N bootstrap distributions to estimate the null distribution. We call the resulting distribution the “aggregated bootstrap” and include it in our numerical studies (see the previous version of the paper for its asymptotic property). For each method and each simulation, N estimated p values \hat{P}_j are obtained. Let N_1 denote the number of p values that are no larger than α_N ; see (5). Then N_1/N is the empirical fraction of null hypotheses that are rejected. When $c = 1$, $N_1/(N\alpha_N) - 1$ reflects the accuracy with which we approximate p values. Its root mean square error (RMSE),

Table 1. RMSEs of $N/(v\alpha) - 1$

α	$n = 6$			$n = 20$			$n = 50$		
	.02	.01	.005	.02	.01	.005	.02	.01	.005
Normal	3.425	5.604	9.083	.833	1.221	1.768	.388	.528	.696
t	.459	.494	.512	.258	.329	.391	.242	.284	.313
Bootstrap	.546	.644	.657	.201	.282	.296	.224	.250	.244
A-bootstrap	.842	.946	.990	.202	.297	.352	.228	.249	.262

NOTE: In model (20), $a_i \equiv .25$ and $b_i \equiv .1$.

$\{E(N_1/(N\alpha_N) - 1)^2\}^{1/2}$, will be reported, where the expectations are approximated by averages across simulations.

We take $N = 600$ (small), $N = 1,800$ (moderate), and $N = 6,000$ (typical) for microarray applications (after preprocessing, which filters out many low-quality measurements on certain genes) and $\alpha_N = 1.5N^{-2/3}$, resulting in $\alpha_N = .02, .01$, and $.005$. The sample size n is taken to be 6 (typical number of microarrays), 20 (moderate), and 50 (large); the case where $n = 6$ is of interest because it is beyond the scope of theory discussed in earlier sections. The number of replications in simulations is $600,000/N$. For the bootstrap calibration method and the aggregated bootstrap, we replicate bootstrap samples 2,000, 4,000, and 9,000 times for $\alpha_N = .02, .01$, and $.005$.

Tables 1 and 2 report the accuracy of estimated p values when $c = 1$. It can be seen that the normal approximations are too inaccurate to be useful. Therefore, we exclude the normal method in the discussion that follows. For $n = 20$ and 50, the bootstrap method provides better approximations than the Student t method. This indicates that the bootstrap can test more hypotheses simultaneously, which is in accordance with our asymptotic theory. Overall the bootstrap method is also slightly better than the aggregated bootstrap, although the two methods are effectively comparable. However, with the small sample size, $n = 6$, the Student t method is relatively the best, although the approximations are poor in general. This is understandable, because the noise distribution is not normal. With such a small sample size, the two bootstrap-based methods, particularly the aggregated bootstrap method, suffer more from random fluctuation in the original samples.

4.2 Real Data Example

The data discussed here were analyzed by Fan et al. (2004), where the biological aim was to examine the impact of the stimulation by MIF, a growth factor, on the expressions of genes in neuroblastoma cells. Six arrays of cDNA microarray data were collected, consisting of relative expression profiles of 19,968 genes in the MIF-stimulated neuroblastoma cells (treatment)

Table 2. RMSEs of $N/(v\alpha) - 1$

α	$n = 6$			$n = 20$			$n = 50$		
	.02	.01	.005	.02	.01	.005	.02	.01	.005
Normal	3.351	5.596	9.014	.770	1.189	1.707	.339	.526	.526
t	.406	.485	.456	.307	.273	.347	.182	.299	.299
Bootstrap	.564	.637	.677	.202	.262	.322	.162	.284	.284
A-bootstrap	.851	.941	.985	.201	.289	.379	.165	.278	.278

NOTE: In model (20), $a_i \sim U(0, .4)$ and $b_i \sim U(0, .2)$.

Table 3. Numbers of genes that are significant at the level α

α	Normal	t	Bootstrap	A-bootstrap
.200	3,205 _(42.26)	2,564 _(33.81)	1,788 _(23.58)	1,565 _(20.64)
.100	2,072 _(27.32)	1,233 _(16.26)	563 _(7.42)	299 _(3.94)
.050	1,328 _(17.51)	504 _(6.65)	143 _(1.89)	31 _(.41)
.010	499 _(6.58)	48 _(.63)	11 _(.15)	1 _(.01)
.005	336 _(4.43)	17 _(.22)	5 _(.07)	0 ₍₀₎
.001	139 _(1.83)	2 _(.03)	0 ₍₀₎	0 ₍₀₎

NOTE: The figures in parentheses are the percentages of significant genes. The total number of genes is 7,583. The bootstrap sampling was repeated 20,000 times.

and those without stimulation (control). After preprocessing that filtered out low-quality expression profiles, 15,266 genes remained. Within-array normalization, discussed by Fan et al. (2004), was applied to remove the intensity effect and block effect. Among the 15,266 gene expression profiles, for simplicity of illustration, we focused only on the 7,583 genes that do not have any missing values. In our notation, $N = 7,583$ and $n = 6$. Table 3 summarizes the results at different levels of significance.

Different methods for estimating p values yield very different results. In particular, the distribution of p values computed by looking up the normal table is stochastically much larger than that based on the t table, which in turn is stochastically much larger than that based on the bootstrap method. The results are very different. As noted earlier, the normal approximation is usually very poor and grossly inflates the number of significant genes. The t approximation and the bootstrap approximation appear more reasonable.

5. PROOFS OF RESULTS IN SECTION 3

For the sake of brevity, we derive only Theorems 1 and 2. Let $C_1 > 0$. Given a random variable X with $E(X) = 0$, consider the condition

$$E(X) = 0, \quad E(X^2) = 1, \quad E(X^4) \leq C_1. \quad (20)$$

The following result follows from theorem 1.2 of Wang (2005), after transforming the distribution of T to that of $(\sum_i X_i)/(\sum_i X_i^2)^{1/2}$.

Theorem 3. Let X, X_1, X_2, \dots denote independent and identically distributed random variables such that (20) holds. Write $T = T(n)$ for the Student t statistic computed from the sample X_1, \dots, X_n , with (for the sake of definiteness) divisor n rather than $n - 1$ used for the variance. Put $\pi_3 = -\frac{1}{3}\kappa_3$, where κ_3 denotes the skewness of the distribution of $X/(\text{var } X)^{1/2}$. Then

$$\frac{P(T > x)}{1 - \Phi(x)} = \exp(\pi_3 x^3 n^{-1/2}) \left\{ 1 + \theta \frac{(1+x)^2}{n^{1/2}} \right\}, \quad (21)$$

where $\theta = \theta(x, n)$ satisfies $|\theta(x, n)| \leq C_2$ uniformly in $0 \leq x \leq C_3 n^{-1/4}$ and $n \geq 1$, and $C_2, C_3 > 0$ depend only on C_1 .

Theorem 1 in the case of normal calibration follows directly from Theorem 3. The case of Student t calibration can be treated similarly.

To derive Theorem 2, note that each $\text{var}(\epsilon'_i) = 1$. To check this, with probability at least $p_n \equiv 1 - \exp(-d_1 n^{1/2})$ for a constant $d_1 > 0$, the conditions of Theorem 3 hold for the bootstrap

distribution of the statistic T_i^* , for each $1 \leq i \leq N$, it suffices to show that there exist constants $0 < C_4 < C_5^{1/2}$ such that, with probability at least p_n , the following condition holds for $1 \leq i \leq N$:

$$C_4 \leq \frac{1}{n} \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2, \quad \frac{1}{n} \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^4 \leq C_5. \quad (22)$$

This can be done using Bernstein's inequality (e.g., Pollard 1984, p. 193) and the assumption that for each i , $P(|\epsilon'_i| \leq C) = 1$. It also can be shown by the uniform convergence result of the empirical process of Korosok and Ma (2007).

Let \mathcal{E}_n denote the event that (22) holds for each $1 \leq i \leq N$. When \mathcal{E}_n prevails, we may apply Theorem 3 to the distribution of T_i^* conditional on \mathcal{Y}_i , obtaining

$$P(T_i^* > x | \mathcal{Y}_i) = \{1 - \Phi(x)\} \exp\left(-\frac{1}{3} \hat{\kappa}_{i3} n^{-1/2} x^3\right) \times \left\{ 1 + \hat{\theta}_i \frac{(1+x)^2}{n^{1/2}} \right\}, \quad (23)$$

where $\hat{\kappa}_{i3}$ is the empirical version of κ_{i3} computed from \mathcal{Y}_i and, in the event that the probability equals $1 - O\{\exp(-d_2 n^{1/2})\}$, $|\hat{\theta}_i| \leq D_1$ uniformly in i and in $0 \leq x \leq x_n$. Here and later, x_n denotes any sequence diverging to infinity but satisfying $x_n = o(n^{1/4})$, and D_1, D_2, \dots and d_1, d_2, \dots denote constants. It follows directly from Theorem 3 that

$$P_{0i}(T_i > x) = \{1 - \Phi(x)\} \exp\left(-\frac{1}{3} \kappa_{i3} n^{-1/2} x^3\right) \times \left\{ 1 + \theta \frac{(1+x)^2}{n^{1/2}} \right\}, \quad (24)$$

where $|\theta_i| \leq D_1$ uniformly in i and in $0 \leq x \leq x_n$.

Result (24), and its analog for the left tail of the distribution of T_i , allow us to express $t_{i\alpha}$, the solution of the equation $P_{0i}(|T_i| > t_{i\alpha}) = 1 - (1 - \alpha)^{1/N}$, as a Taylor expansion,

$$|t_{i\alpha} - z_\alpha - c \kappa_{i3} n^{-1/2} z_\alpha^2| \leq D_2 (n^{-1} z_\alpha^4 + n^{-1/2} z_\alpha),$$

uniformly in i , where c is a constant and z_α is the solution of $P(|Z| > z_\alpha) = 1 - (1 - \alpha)^{1/N}$. Note that if z_α solves this equation, then $z_\alpha \sim (2 \log N)^{1/2}$, and so, because $\log N = o(n^{1/2})$, $z_\alpha = o(n^{1/4})$. Therefore, without loss of generality, $0 \leq z_\alpha \leq x_n$. Likewise, we may assume that $0 \leq t_{i\alpha} \leq x_n$ and $0 \leq \hat{t}_{i\alpha} \leq x_n$ with probability $1 - O\{\exp(-d_2 n^{1/2})\}$.

Also, from (23), we can see that in the event that the probability equals $1 - O\{\exp(-d_2 n^{1/2})\}$,

$$|\hat{t}_{i\alpha} - z_\alpha - c \hat{\kappa}_{i3} n^{-1/2} z_\alpha^2| \leq D_3 (n^{-1} z_\alpha^4 + n^{-1/2} z_\alpha).$$

However, in the event that probability equals that $1 - O\{\exp(-d_3 n^{1/2})\}$, $|\hat{\kappa}_{i3} - \kappa_{i3}| \leq D_4 n^{-1/4}$, and therefore, in the event with probability equals $1 - O\{\exp(-d_4 n^{1/2})\}$,

$$|\hat{t}_{i\alpha} - z_\alpha - c \kappa_{i3} n^{-1/2} z_\alpha^2| \leq D_5 (n^{-1} z_\alpha^4 + n^{-1/2} z_\alpha + n^{-3/4} z_\alpha^2).$$

It follows from the foregoing results that $P_{0i}(|T_i| > \hat{t}_{i\alpha})$ lies between the respective values of

$$P_{0i}(|T_i| > t_{i\alpha} \pm \delta) \mp D_6 \exp(-d_4 n^{1/2}), \quad (25)$$

where

$$\delta = D_5 (n^{-1} z_\alpha^4 + n^{-1/2} z_\alpha + n^{-3/4} z_\alpha^2).$$

Using (24) and its analog for the left tail to expand the probability in (25), we can deduce that

$$P_{0i}(|T_i| > t_{i\alpha} \pm \delta) = P_{0i}(|T_i| > t_{i\alpha})\{1 + o(1)\}$$

uniformly in i . More simply, $\exp(-d_4 n^{1/2}) = o\{P_{0i}(|T_i| > t_{i\alpha})\}$, using the fact that $z_\alpha = o(n^{1/4})$ and $\exp(-D_7 z_\alpha^2) = o\{P_{0i}(|T_i| > t_{i\alpha})\}$ for sufficiently large $D_7 > 0$. Thus

$$P_{0i}(|T_i| > \hat{t}_{i\alpha}) = P_{0i}(|T_i| > t_{i\alpha})\{1 + o(1)\};$$

uniformly in i . Theorem 2 follows from this property.

[Received April 2006. Revised July 2007.]

REFERENCES

- Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57, 289–300.
- Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188.
- Bentkus, V., and Götze, F. (1996), "The Berry–Esseen Bound for Student's T-statistic," *The Annals of Statistics*, 24, 491–503.
- Bickel, P. J., and Levina, E. (2004), "Some Theory for Fisher's Linear Discriminant, 'Naive Bayes,' and Some Alternatives When There Are Many More Variables Than Observations," *Bernoulli*, 10, 989–1010.
- Clarke, S., and Hall, P. (2007), "Robustness of Multiple Testing Procedures Against Independence," unpublished manuscript.
- Donoho, D. L., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *The Annals of Statistics*, 32, 962–994.
- Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103.
- Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104.
- (2007), "Correlation and Large-Scale Simultaneous Significance Testing," *Journal of the American Statistical Association*, 102, 93–103.
- Fan, J., and Li, R. (2006), "Statistical Challenges With High Dimensionality: Feature Selection in Knowledge Discovery," in *Proceedings of the International Congress of Mathematicians*, Vol. III, eds. M. Sanz-Sole, J. Soria, J. L. Varona, and J. Verdera, Zurich: European Mathematical Society, pp. 595–622.
- Fan, J., Chen, Y., Chan, H. M., Tam, P., and Ren, Y. (2005a), "Removing Intensity Effects and Identifying Significant Genes for Affymetrix Arrays in MIF-Uppressed Neuroblastoma Cells," *Proceeding of the National Academy of Science USA*, 102, 17751–17756.
- Fan, J., Peng, H., and Huang, T. (2005b), "Semilinear High-Dimensional Model for Normalization of Microarray Data: A Theoretical Analysis and Partial Consistency" (with discussion), *Journal of the American Statistical Association*, 100, 781–813.
- Fan, J., Tam, P., Vande Woude, G., and Ren, Y. (2004), "Normalization and Analysis of cDNA Micro-Arrays Using Within-Array Replications Applied to Neuroblastoma Cell Response to a Cytokine," *Proceedings of the National Academy of Science USA*, 101, 1135–1140.
- Finner, H., and Roters, M. (1998), "Asymptotic Comparison of Step-Down and Step-Up Multiple Test Procedures Based on Exchangeable Test Statistics," *The Annals of Statistics*, 26, 505–524.
- (1999), "Asymptotic Comparison of the Critical Values of Step-Down and Step-Up Multiple Comparison Procedures," *Journal of Statistical Planning and Inference*, 79, 11–30.
- (2000), "On the Critical Value Behaviour of Multiple Decision Procedures," *Scandinavian Journal of Statistics*, 27, 563–573.
- (2001), "On the False Discovery Rate and Expected Type I Errors," *Biometrical Journal*, 43, 985–1005.
- (2002), "Multiple Hypotheses Testing and Expected Number of Type I Errors," *The Annals of Statistics*, 30, 220–238.
- Genovese, C., and Wasserman, L. (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061.
- Hall, P. (1990), "On the Relative Performance of Bootstrap and Edgeworth Approximations of a Distribution Function," *Journal of Multivariate Analysis*, 35, 108–129.
- Hall, P., and Wang, Q. (2007), "Strong Approximations of Level Exceedences Related to Multiple Hypothesis Testing," unpublished manuscript.
- Hu, J., and He, X. (2007), "Enhanced Quantile Normalization of Microarray Data to Reduce Loss of Information in the Gene Expression Profile," *Biometrics*, 63, 50–59.
- Huang, J., Wang, D., and Zhang, C. (2005), "A Two-Way Semi-Linear Model for Normalization and Significant Analysis of cDNA Microarray Data," *Journal of the American Statistical Association*, 100, 814–829.
- Korosok, M. R., and Ma, S. (2007), "Marginal Asymptotics for the 'Large p , Small n ' Paradigm: With Applications to Microarray Data," *The Annals of Statistics*, 35, 1458–1486.
- Lehmann, E. L., and Romano, J. P. (2005), "Generalizations of the Familywise Error Rate," *The Annals of Statistics*, 33, 1138–1154.
- Lehmann, E. L., Romano, J. P., and Shaffer, J. P. (2005), "On Optimality of Stepdown and Stepup Multiple Test Procedures," *The Annals of Statistics*, 33, 1084–1108.
- Pollard, D. (1984), *Convergence of Stochastic Processes*, New York: Springer.
- Reiner, A., Yekutieli, D., and Benjamini, Y. (2003), "Identifying Differentially Expressed Genes Using False Discovery Rate Controlling Procedures," *Bioinformatics*, 19, 368–375.
- Sarkar, S. K. (2002), "Some Results on False Discovery Rate in Stepwise Multiple Testing Procedures," *The Annals of Statistics*, 30, 239–257.
- (2006), "False Discovery and False Nondiscovery Rates in Single-Step Multiple Testing Procedures," *The Annals of Statistics*, 34, 394–415.
- Shao, Q.-M. (1999), "A Cramér-Type Large Deviation Result for Student's t -Statistic," *Journal of Theoretic Probability*, 12, 385–398.
- Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society, Ser. B*, 66, 187–205.
- van der Laan, M. J., and Bryan, J. (2001), "Gene Expression Analysis With the Parametric Bootstrap," *Biostatistics*, 2, 445–461.
- Wang, Q. (2005), "Limit Theorems for Self-Normalized Large Deviations," *Electronic Journal of Probability*, 10, 1260–1285.