TO HOW MANY SIMULTANEOUS HYPOTHESIS TESTS CAN NORMAL, STUDENT'S t OR BOOTSTRAP CALIBRATION BE APPLIED?

Jianqing Fan Peter Hall Qiwei Yao

ABSTRACT. 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, ν 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 for combining data across samples. In this setting, how large can ν be, as a function of n, before level accuracy becomes poor? In the present paper we answer this question in cases where the statistic under test is of Student's t type. We show that if either normal or Student's t distribution is used for calibration then the level of the simultaneous test is accurate provided $\log \nu$ increases at a strictly slower rate than $n^{1/3}$ as n diverges. If $\log \nu$ and $n^{1/3}$ diverge at the same rate then asymptotic level accuracy requires the average value of standardised skewness, taken over all distributions to which the tests are applied, to converge to zero as n increases. On the other hand, if bootstrap methods are used for calibration then significantly larger values of ν are feasible; we may choose $\log \nu$ almost as large as $n^{1/2}$ and still achieve asymptotic level accuracy, regardless of the values of standardised skewness. It seems likely that similar conclusions hold for statistics more general than the Studentised mean, and that the upper bound of $n^{1/2}$, in the case of bootstrap calibration, can be increased.

KEYWORDS. Bonferroni's inequality, Edgeworth expansion, genetic data, large-deviation expansion, level accuracy, microarray data, quantile estimation, skewness, Student's t statistic.

Jianqing Fan is Professor (E-mail: jqfan@princeton.edu), Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. Peter Hall (E-mail: halpstat@maths.anu.edu.au) is professor, Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia. Qiwei Yao is professor (E-mail: q.yao@lse.ac.uk), Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom.
Fan's work was sponsored in part by NSF grant DMS-0354223 and NIH grant R01-GM07261.

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, ν, is frequently much larger than sample size, n. Indeed, sample size is typically small; e.g. $n = 8$, 20 or 50 are considered to be typical, moderately large or large, respectively, for microarray data. The question arises naturally as to how large ν can be before the accuracy of simultaneous statistical inference becomes poor.

An important study in this direction was initiated by van der Laan and Bryan (2001). These authors showed that the population mean and variance parameters can be consistently estimated when $(\log \nu)/n \to 0$ if observed data are bounded. See also Bickel and Levina (2004) for a similar result in a high-dimensional classification problem. In the context of the normalization of microarray data, Fan, Peng and Huang (2005) and Huang, Wang and Zhang (2005) studied semiparametric problems in a framework where ν tended to ∞, and discovered new and deep insights into semiparametric problems. Hu and He (2006) proposed an enhanced quantile normalization based on the high-dimensional singular value decomposition to reduce information loss in gene expression profiles. See also He and Wang (2006) for recent development and references on estimation and testing problems arising from analysis of Affymetrix arrays. The high-dimensional statistical inference problems have been studied in the pioneering work by Huber (1973) and Portnoy (1988).

Recently, Korosok and Ma (2005) significantly widened the spectrum of simultaneous inference by first discussing simultaneous convergence of the marginal empirical distribution $\hat{F}_{n,i}$, based on measurements on the ith gene, to its theoretical counterpart, F_i. They demonstrated convincingly that under suitable conditions,

$$\max_{1 \le i \le \nu} \|\hat{F}_{n,i} - F_i\|_\infty \to 0\,,$$

when $(\log \nu)/n \to 0$. As a corollary, they showed that the approximated P-value $\hat{p}_i$ of a t-type test statistic for testing whether the ith marginal mean is zero, computed under the normal approximation, converges uniformly to its asymptotic counterpart p_i, i.e.

$$\max_{1 \le i \le \nu} \|\hat{p}_i - p_i\|_\infty \to 0\,, \tag{1.1}$$

when $\log \nu = o(n^{1/2})$ and the observed data fall into a bounded interval. In addition, they showed that the uniform approximation holds for median based tests when $\log \nu = o(n^{1/3})$. These results are important advances in the literature of simultaneous testing, where P-values are popularly assumed to be known. See, for example, Benjamini and Yekutieli (2001), Dudoit, Shaffer and Boldrick (2003), Donoho and Jin (2004), Efron (2004), Genovese and Wasserman (2001), Storey, Taylor and Siegmund (2004), Lehmann and Romano (2005), Lehmann, Romano and Shaffer (2005) where many new ideas have been introduced to control different aspects of false discovery rates.

However, the fundamental assumption that the P-values are calculated without error is unrealistic in practice. The question then arises as to which approximation methods are more accurate in calculating P-values, and how many of those values can be approximated simultaneously. The approximation (1.1) is not adequate for multiple comparison problems.

Take the celebrated Benjamini and Hochberg (1995) method as an example. It is assumed that we have ν P-values $p_1, \ldots, p_\nu$ available for testing the marginal hypotheses. If the false discovery rate (FDR) is controlled at p, then k_n hypotheses with the smallest P-values are rejected, where

$$k_n = \max\{i : p_{(i)} \le ip/\nu\} \tag{1.2}$$

and $p_{(i)}$ denotes the ith smallest P-value: $p_{(1)} \le p_{(2)} \le \ldots \le p_{(\nu)}$. If the P-values need to be estimated, the accuracy of the estimators should be of order $o(1/\nu)$. This means that we need to approximate the extreme tail probability under the null hypotheses. For this purpose, the approximation (1.1) requires significant refinement.

The requirement above is relaxed by Fan *et al.* (2004). In the analysis of gene expression data, Fan *et al.* (2004) take $\alpha_\nu = 0.001$ and find the significant set of genes

$$\mathcal{S} = \{j : p_j \leq \alpha_\nu, j = 1, \ldots, \nu\} \tag{1.3}$$

for $\nu = 15{,}000$ simultaneous tests. Note that α_ν is an order of magnitude larger than ν^{-1}. Hence, the approximation errors allowed for estimating P-values are less stringent for computing (1.3) than the order ν^{-1}. However, we still need to approximate the tail probability, namely, $\alpha_\nu \to 0$ and $\nu\alpha_\nu \to \infty$. Lehmann and Romano (2005) demonstrate that the probability that the falsely is more than α_1' is no more than α_2' as long as than α_2' as long as $\alpha_1'\alpha_2' = \alpha_\nu$. For example, the probability that the proportion of genes falsely discovered in (1.3) exceeds 10%, is no more than 1%.

The number of elements in $\mathcal{S}$, denoted by k_n', equals the number of genes discovered at the significance level α_ν. Note that $\nu\alpha_\nu$ is the upper bound to the expected number of falsely discovered genes and is approximately the same as the expected number of falsely discovered genes when most null hypotheses are true. Hence, the FDR is estimated as $\hat{p} = \nu\alpha_\nu/k_n'$, which gives us an idea of the quality of the selected genes; see Fan *et al.* (2004). This simple procedure is closely related to the Benjamini and Hochberg (1995) procedure for controlling the false discovery rate. More precisely, let $\hat{k}_n$ be given by (1.2) when p is taken as the estimated FDR, $\hat{p}$, namely,

$$\hat{k}_n = \max\{i : p_{(i)} \leq i\hat{p}/\nu\}. \tag{1.4}$$

According to Fan *et al.* (2005), $k_n' \leq \hat{k}_n$, but these quantities are often very close. See Table 3 in section 5. Note that when $k_n' = \hat{k}_n$, both the Benjamini and Hochberg method and the empirical method of Fan *et al.* (2004) select exactly the same set of genes. Namely, the Benjamini and Hochback method with $p = \hat{p}$ would call results in $\mathcal{S}$ statistically significant genes.

In this paper, we investigate the question of how many hypotheses about the

mean can be tested simultaneously by using Student's t-type of statistic, before the level of aggregated errors becomes poor. We show that if either the normal distribution or the t-distribution is used for calibration, the level of the simultaneous test is accurate when $\log \nu = o(n^{1/3})$ or when $\log \nu = O(n^{1/3})$ and the average of the standardised skewness, taken over all distributions to which the tests are applied, converges to zero. On the other hand, if bootstrap methods are used for estimating P-values, then significantly larger values of ν are allowed. The asymptotic level of the simultaneous test is accurate as long as $\log \nu = o(n^{1/2})$. Thus, the advantages offered by bootstrap calibration are clear. One interesting aspect of our results is that, provided a Bonferroni argument is used to bound simultaneous coverage levels, the dependence structure among ν-dimensional vectors can be arbitrary.

The paper is organized as follows. In section 2, we formulate the accuracy problem for simultaneous tests. There, we also outline statistical models and testing procedures. Our main results are presented in section 3, where we answer the question of how many hypotheses can be tested simultaneously. Section 4 outlines the idea of marginal aggregation when the number of hypotheses is ultra-large. Numerical investigations among various calibration methods are presented in section 5. Technical proofs of results in section 3 are relegated to section 6.

2. MODEL AND METHODS FOR TESTING

2.1. Basic model and methodology.

The simplest model is that where we observe random variables

$$Y_{ij} = \mu_i + \epsilon_{ij}\,, \quad 1 \le i < \infty\,, \quad 1 \le j \le n\,, \tag{2.1}$$

with the index i denoting the ith gene, j indicating the jth array, and the constant μ_i representing the mean effect for the ith gene. We shall assume that:

for each i, $\epsilon_{i1}, \ldots, \epsilon_{in}$ are independent and identically distributed random variables with zero expected value. (2.2)

The discussion and results below are readily extended to the case where $n = n_i$ depends on i, but taking n fixed simplifies our discussion. In practice the number of values of i, at (2.1), would of course be finite rather than infinite, and we shall consider only simultaneous tests for a finite number, ν say, of genes. We take the potential number of genes to be infinite, rather than have a finite value, say m, only so as to avoid having to state repeatedly that $\nu \le m$.

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

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

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 $1 \le i \le \nu$ say. We first study this classical testing problem of controlling the probability of making at least one false discoveries, which requires calculating P-values of accuracy $o(\nu^{-1})$, the same as that needed in (1.2). We then extend the results to control relaxed FDR in (1.3), which is less stringent.

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

$$P(|N| > t_\alpha) = 1 - (1-\alpha)^{1/\nu}\,, \quad P\{|T(n-1)| > t_\alpha\} = 1 - (1-\alpha)^{1/\nu}\,, \tag{2.3}$$

where N and $T(k)$ have respectively the standard normal distribution and the Student's t distribution with k degrees of freedom. We expect that, if t_α satisfies either of the equations at (2.3), and if H_{0i} is true for each i in the range $1 \le i \le \nu$, then the probability that, for at least one of these i's, the test of H_{0i} against H_{1i} leads to rejection, is close to the nominal significance level, α. A key question is: How large can ν be and level accuracy still be good?

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

$$p_i = P_{0i}(|T_i| > t_\alpha)\,, \tag{2.4}$$

where P_{0i} denotes probability calculated under H_{0i}. The question arises as to how large ν can be so that

$$\max_{1\le i\le\nu} p_i = o(1) \quad \text{and} \quad \sum_{i=1}^{\nu} p_i = \beta + o(1)\,, \tag{2.5}$$

for some $0 < \beta < \infty$. This implies that the significance level of the simultaneous test, described in section 2.1, is

$$\alpha(\nu) \equiv P\Big(H_{0i} \text{ rejected for at least one } i \text{ in the range } 1 \le i \le \nu\Big) \tag{2.6}$$

$$\le \sum_{i=1}^{\nu} p_i = \beta + o(1)\,. \tag{2.7}$$

If, in addition to (2.2), we assume that

$$\text{the sets of variables } \{\epsilon_{ij},\ 1 \le j \le n\} \text{ are independent for different } i, \tag{2.8}$$

then

$$\alpha(\nu) = 1 - \prod_{i=1}^{\nu}(1-p_i) = 1 - \exp\Big(-\sum_{i=1}^{\nu} p_i\Big) + O\Big(\sum_{i=1}^{\nu} p_i^2\Big). \tag{2.9}$$

Consequently, (2.5) and (2.9) imply the following property:

$$\text{if (2.8) holds then} \quad \alpha(\nu) = 1 - e^{-\beta} + o(1), \tag{2.10}$$

where $\alpha(\nu)$ is as defined at (2.6). The "$o(1)$" terms in (2.7) and (2.10) are quantities which converge to zero as $\nu \to \infty$.

In practice we would take $\beta = -\log(1-\alpha)$, if we were prepared to assume (2.8) and wished to construct a simultaneous test with level close to α; and take $\beta = \alpha$, if we were using Bonferroni's bound to construct a conservative simultaneous test with the same approximate level.

2.3. Methods for calibration.

For calibration against normal or Student's t distributions we take the critical point t_α to be the solution of the respective equations (2.3). A variety of methods, including explicit Edgeworth correction, can also be used to effect calibration.

Results of Hall (1990) indicate that the bootstrap has advantages over Edgeworth correction in the present setting, where large deviations are involved, so we shall consider the bootstrap.

Let $Y_{i1}^\dagger, \ldots, Y_{in}^\dagger$ denote a bootstrap resample drawn by sampling randomly, with replacement, from $\mathcal{Y}_i = \{Y_{i1}, \ldots, Y_{in}\}$. Put $Y_{ij}^* = Y_{ij}^\dagger - \bar{Y}_i$ and $T_i^* = n^{1/2}\bar{Y}_i^*/S_i^*$, where $\bar{Y}_i^* = n^{-1}\sum_j Y_{ij}^*$ and $(S_i^*)^2 = n^{-1}\sum_j (Y_{ij}^* - \bar{Y}_i^*)^2$. Write z_α for the conventional normal critical point for ν simultaneous tests. That is, z_α solves the equation $P(|N| > z_\alpha) = 1 - (1-\alpha)^{1/\nu}$. (We could also use the Student's t point.) Define $a = \hat{f}_i(\alpha)$ to be the solution of the equation

$$P(|T_i^*| > z_a \mid \mathcal{Y}_i) = 1 - (1-\alpha)^{1/\nu}\,.$$

Equivalently, $\hat{f}_i(\alpha)$ is the bootstrap estimator of the ideal, but unavailable, nominal significance level we would employ in place of α if our aim was to use normal calibration to achieve a test with level exactly equal to α.

Our bootstrap critical point is $\hat{t}_{i\alpha} = z_{\hat{f}_i(\alpha)}$. That is, $\hat{t}_{i\alpha}$ plays the role that t_α did in sections 2.1 and 2.2; we reject H_{0i} if and only if $|T_i| > \hat{t}_{i\alpha}$. Since the critical point is now a random variable, and depends on data in the ith "row" $\mathcal{Y}_i$, although not on data from other rows, then the earlier calculations should be revisited. In particular, the definition of p_i at (2.4) should be replaced by

$$p_i = P_{0i}(|T_i| > \hat{t}_{i\alpha})\,. \tag{2.11}$$

With this new definition, (2.10) continues to be a consequence of (2.5).

2.4. Accuracy of approximations for controlling FDR.

The requirement (2.5) aims at controlling the probability of making at least one falsely discovered genes. For gene screening, it is more appropriate to control the proportion of falsely discovered genes. This is much less stringent than (2.5). Expressions (2.3) and (2.5) require basically the approximation errors

$$p_i = \alpha_\nu\{1 + o(1)\}, \qquad \alpha_\nu = \beta/\nu + o(\nu^{-1}). \tag{2.12}$$

The required accuracy of order $o(1/\nu)$ reflects that needed in the Benjamini-Hochberg method (1.2).

For controlling FDR using (1.3), however, the mathematical requirement is more relaxed than for controlling the cumulative approximation error (2.5). If we take $\alpha_\nu = 1.5\nu^{-2/3}$, say, then the expected number of falsely discovered genes is bounded by $1.5\nu^{1/3}$, or 13, 18, 27, 34, 37 for $\nu = 600, 1800, 6{,}000, 12{,}000, 15{,}000$, respectively. In this case, (1.3) requires maximum approximation error of order

$$\max_{1\le i\le\nu} |p_i - \alpha_\nu| = o(\alpha_\nu)\,,$$

with $\alpha_\nu = 1.5\nu^{-2/3}$. This is a much less stringent requirement than (2.12) and will be covered by our theoretical results in section 3.

3. THEORETICAL RESULTS

3.1. Asymptotic results.

In Theorem 3.1 and Corollary 3.1 below we shall show that normal and Student's t calibration give asymptotically correct calibration, in the sense that (2.5) and hence (2.10) hold for a knowable value of β (not depending on the unknown distributions of ϵ_{i1}), in a general sense (in particular, for skew distributions), if and only if $\log\nu = o(n^{1/3})$. Furthermore, if $\log\nu$ is asymptotic to a constant multiple of $n^{1/3}$ then (2.5) and (2.10) are valid if and only if the limiting average value of absolute skewnesses of the first ν distributions of ϵ_{i1} equals zero.

On the other hand, if we use bootstrap calibration, and even if all the error distributions are skew, the asymptotically correct level for a simultaneous test is achieved with $\log\nu$ as large as $o(n^{1/2})$. See Theorem 3.3. This result shows one of the advantages of bootstrap calibration, compared with calibration using Student's t and normal distributions: Bootstrap calibration allows us to apply a larger number of simultaneous tests before level-accuracy problems arise.

Define κ_{i3} to be the third cumulant, or equivalently the skewness, of the distribution of $\epsilon_i' = \epsilon_{i1}/(E\epsilon_{i1}^2)^{1/2}$.

Theorem 3.1. *Assume that*

$$\max_{1\le i\le\nu} E|\epsilon_i'|^3 = O(1) \tag{3.1}$$

as $\nu\to\infty$, *and suppose too that* $\nu=\nu(n)\to\infty$ *in such a manner that* $(\log\nu)/n^{1/3}\to\gamma$, *where* $0\le\gamma<\infty$. *Define* t_α *by either of the formulae at* (2.3), *and* p_i *by* (2.4). *Then* (2.5) *holds with*

$$\beta=\beta(\nu)\equiv-\frac{\log(1-\alpha)}{\nu}\sum_{i=1}^{\nu}\cosh\left(\tfrac{1}{3}\,\gamma^3\kappa_{i3}\right), \tag{3.2}$$

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

The value of $\beta(\nu)$, defined at (3.2), is bounded by $|\log(1-\alpha)|\cosh(\gamma^3 B)$, uniformly in ν, where $B=\sup_i|\kappa_{i3}|$.

Corollary 3.2. *Assume the conditions of Theorem 3.1. If* $\gamma=0$, *i.e. if* $\log\nu=o(n^{1/3})$, *then* (2.5) *and* (2.10) *hold with* $\beta=-\log(1-\alpha)$; *and if* $\gamma>0$ *then* (2.5) *holds with* $\beta=-\log(1-\alpha)$ *if and only if* $\nu^{-1}\sum_{i\le\nu}|\kappa_{i3}|\to0$, *i.e. if and only if the limit of the average absolute values of the skewnesses of the distributions of* $\epsilon_{11},\ldots,\epsilon_{\nu1}$ *equals zero.*

Theorem 3.3. *Strengthen* (3.1) *to the assumption that for a constant* $C>0$, $P(|\epsilon_i'|\le C)=1$, *and suppose too that* $\nu=\nu(n)\to\infty$ *in such a manner that* $\log\nu=o(n^{1/2})$. *Define* $\hat t_{i\alpha}=t_{\hat f_i(\alpha)}$, *as in section 2.4, and define* p_i *by* (2.11). *Then* (2.5) *holds with* $\beta=-\log(1-\alpha)$.

3.2. Applications to controlling FDR.

Our proofs in section 6 show that, under the conditions of Theorem 3.1,

$$P_{0i}(|T_i|>t_{i\alpha})=\beta/\nu+o(\nu^{-1})$$

with $\beta=-\log(1-\alpha)$, uniformly in i under the null hypotheses, when $\log\nu=o(n^{1/3})$ or $\log\nu=O(n^{1/3})$ with the additional assumption that the skewness satisfies the

condition of Corollary 3.2. In addition, under the conditions of Theorem 3.3,

$$P_{0i}(|T_i| > \hat{t}_{i\alpha}) = \beta/\nu + o(\nu^{-1})$$

uniformly in i under the null hypotheses when $\log \nu = o(n^{1/2})$. These improve a uniform convergence result given by Kosorok and Ma (2005), at the expense of more restrictions on ν.

When the P-values in (1.2) need to be estimated, the estimation errors should be of order $o(\nu^{-1})$, where ν diverges with n. On the other hand, when P-values in (1.3) are estimated, the precision can be of order $o(\alpha_\nu)$, where $\alpha_\nu = b_\nu/\nu$ with $b_\nu \to \infty$. In this case, the large deviation results in Theorems 3.1 and 3.3 continue to be applicable.

Note that the tail probability of the standard normal distribution $P(|N| \geq x_n)$ is of order $\exp(-x_n^2/2)/x_n$. Suppose that the large deviation result holds up to the point x_n, which is of order $o(n^{1/3})$ for Student's t calibration and $o(n^{1/4})$ for bootstrap calibration (see section 6). Setting it equal to α_ν yields

$$\log \nu - \log b_\nu = x_n^2/2 + \log x_n. \tag{3.3}$$

This puts a limit on the number of simultaneous tests that can be performed with a good approximation of P-values. A larger b_ν allows a larger value of ν. For example, if we take $b_\nu = 1.5/\nu^{1/3}$ as in section 2.4, then

$$\log \nu = 3x_n^2/4 + 1.5 \log x_n + \log 1.5,$$

which is much larger than that in the case where $b_\nu = 1$ is used in (1.2) or (2.5). To put this into perspective, let ν_1 and ν_2 be the numbers of simultaneous hypotheses allowed with $b_\nu = 1$ and $b_\nu = 1.5\nu^{1/3}$, respectively. Namely, they solve (3.3) respectively with $b_\nu = 1$ and $b_\nu = 1.5\nu^{1/3}$. Then, $\nu_2 = 1.5\nu_1^{3/2}$. For example, if $\nu_1 = 500$ then $\nu_2 = 16{,}771$.

For bootstrap calibration, when $x_n = o(n^{1/4})$, by the proof of Theorem 3.2 we have

$$\max_{1\le i\le \nu} |P_{0i}(|T_i| > \hat{t}_{i\alpha})/P(|N| > x_n) - 1| = o(1),$$

where $\alpha = P(|N| > x_n)$. Substituting $x_n = o(n^{1/4})$ into (3.3), $\nu = \exp\{o(n^{1/2})\}$. The larger b_ν implies a larger constant factor in $o(n^{1/2})$ or x_n^2, as illustrated in the last paragraph. On the other hand, for Student's t calibration, $\nu = \exp\{o(n^{1/3})\}$.

4. MARGINAL AGGREGATION

4.1. Methods of Aggregation.

We have demonstrated that with bootstrap calibration, Student's t-statistics can test simultaneously a number of hypotheses of order $\exp\{o(n^{1/2})\}$. Although this value may be conservative, it may still not be large enough for some applications to microarray and tiling arrays where the number of simultaneous tests can be even larger. Similarly, whether looking up a t-table and normal table depends very much on mathematical assumptions. For example, suppose an observed value of a t statistic is 6.7. Its corresponding two-tail P-value, for $n = 6$ arrays, is 0.112% when looking up t-tables with five degrees of freedom, and 2.084×10^{-11} when consulting normal tables. If, as mentioned in the introduction, $\nu = 15{,}000$, then using criterion (1.3) with $\alpha = 0.1\%$, the null hypothesis with the observed t-statistic 6.7 will not be rejected. The P-value calculation depends here heavily on mathematical assumptions, and is not robust to the error distribution. On the other hand, when $n = 8$ the situation eases dramatically; the P-value is now 0.0277% under the t-distribution with degrees of freedom 7, for an observed t-statistic value of 6.7.

To overcome the afforementioned problems, Reiner, Yekutieli and Benjamini (2003) and Fan *et al.* (2004) introduce marginal aggregation methods. The basic assumption is that the null distributions of test statistics T_i are the same, denoted by F. With this assumption, we can use the empirical distribution of $\{T_i, i =$

$1, \ldots, \nu\}$, i.e.

$$\hat{F}_\nu(x) = \nu^{-1} \sum_{i=1}^{\nu} I(T_i \le x)\,, \tag{4.1}$$

as an estimator of F. This turns the "curse-of-dimensionality" into a "blessing-of-dimensionality". In this case, even if n is finite, the distribution of F can still be consistently estimated when $\nu \to \infty$ and only a small fraction of alternative hypotheses are true.

Reiner, Yekutieli and Benjamini (2003) and Fan *et al.* (2004) carry this idea one step further. They aggregate the estimated distributions for each T_i, based on a resampling technique (more precisely, a permutation method). For example, we can use the average of bootstrap estimators,

$$\hat{F}^*_\nu(x) = \nu^{-1} \sum_{i=1}^{\nu} P(T'_i < x \mid \mathcal{Y}_i)\,, \tag{4.2}$$

as an estimator of $F(x)$. In implementation, we draw B bootstrap samples and compute B bootstrap statistics for each marginal distribution, resulting in

$$\{T^*_{ij}, i = 1, \ldots, \nu, j = 1, \ldots, B\}\,. \tag{4.3}$$

The aggregated estimator $\hat{F}^*_\nu(x)$ can be approximated by the empirical distribution for the pseudo-data in (4.3).

Fan *et al.* (2005) propose a sieve idea for reducing the possible bias of $\hat{F}$ and $\hat{F}^*$. The basic idea is to pre-screen the sets of hypotheses that might be statistically significant, resulting in a subset, $\mathcal{I}$, which is not statistically significant (this can be done by using the normal approximation with a relaxed P-value), and then restricting the aggregations in (4.1) or (4.2) to subset $\mathcal{I}$. Note that, under model (2.1), for the bootstrap t-statistic, $E\{P(T^*_i < x \mid \mathcal{Y}_i)\}$ does not depend on the unknown value μ_i and there is no need for the sieve method when the bootstrap method is used. The asymptotic theory of this kind of aggregated bootstrap is poorly understood. We shall investigate its large sample properties in the next section, in the context of approximating tail probabilities.

4.2. Asymptotic theory.

The asymptotic theory requires the assumption that the variables $\{T_i\}$ are only very weakly dependent. In particular, we assume that $\{T_i\}$ are pairwise nearly tail-independent. This assumption holds under condition (2.8). In the following statement, x_n denotes the left-tail critical value.

Theorem 4.1. *Let ν_1 be the number of nonzero μ_i, i.e. the number of elements in $\{i : H_{1i}$ is true$\}$. Then,*

$$\hat{F}_\nu(x_n) = F(x_n) + O_P\left[(\nu_1/\nu) + \sqrt{F(x_n)/\nu}\left\{1 + a_n\nu + a_n\,\nu_1\,F(x_n)^{-1/2}\right\}^{1/2}\right],$$

provided that $|r_{ij}| \le a_n$, with r_{ij} denoting the correlation coefficient between $I(T_i < x_n)$ and $I(T_j < x_n)$.

A similar result holds for the upper tail. Note that the term $O(\nu_1/\nu)$ reflects the bias of the estimate. It can be reduced by using the sieve idea of Fan *et al.* (2005). Further discussion on this topic is beyond the scope of our paper. When we estimate the P-value at order $F(x_n) = b_\nu/\nu$ with $b_\nu \to \infty$ and $b_n/\nu \to 0$, the approximation error in Theorem 4.1 is of order $o_P\{F(x_n)\}$, if $\nu_1 = o(b_\nu)$ and $\nu a_n = o(b_\nu)$. This approximation is accurate enough for applications using criterion (1.3) for controlling FDR or selecting significant hypotheses.

We now describe the asymptotic result for the aggregated bootstrap distribution. To reduce technicalities we assume that (2.8) holds and $\{\epsilon_{ij}\}$ for different i are identically distributed up to a scale transform. Then, we have:

Theorem 4.2. *Under the assumptions above,*

$$\hat{F}^*_\nu(x_n) = F_n(x_n) + O_P\left\{\sqrt{F_n(x_n)/\nu}\right\},$$

where $F_n(x) = E\{P(T^*_i < x \mid \mathcal{Y}_i)\}$.

As mentioned before, $F_n(x)$ does not depend on unknown parameters μ_i. It admits large deviation expansions similar to (6.4). Thus, it has approximately the

same accuracy as the marginal bootstrap method. However, it is generally a biased estimator of the null distribution $F(x)$.

5. NUMERICAL PROPERTIES

We have presented five methods for computing P-values: the normal, t and bootstrap calibrations and the two marginal aggregation methods (4.1) and (4.2). Our theoretical results indicate that the bootstrap calibration method enables us to simultaneously test the number of hypotheses up to at least $\exp\{o(n^{1/2})\}$, better than the simple normal and t calibrations. Under the marginal-distribution assumption that the null distributions are identical and the true alternative hypotheses are sparse, the empirical method (4.1) allows us to test even larger values of ν. The "aggregated bootstrap" method (4.2) has a performance similar to that of the bootstrap calibration method.

In our simulation study we attempt to construct the models that reflect some aspects of gene expression data. To this end, we divide genes into three groups. Within each group, genes share one unobserved common factor with different factor loadings. In addition, there is an unobserved common factor among all the genes across the three groups. For simplicity of presentation, we assume that ν is a multiple of three. We denote by $\{Z_{ij}\}$ a sequence of independent $\mathrm{N}(0,1)$ random variables, and $\{\chi_{ij}\}$ a sequence of independent random variables of the same distribution as that of $(\chi^2_m - m)/\sqrt{2m}$. 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 (2.1) is defined as

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

where $a_{ij} = 0$ except that $a_{i1} = a_i$ for $i = 1, \ldots, \nu/3$, $a_{i2} = a_i$ for $i = \nu/3 + 1, \ldots, 2\nu/3$, and $a_{i3} = a_i$ with $i = (2\nu/3) + 1, \ldots, \nu$. Note that $E\epsilon_{ij} = 0$ and $\mathrm{var}(\epsilon_{ij}) = 1$, and that the within-group correlation is in general stronger than the

between-group correlation, since the former shares one extra common factor. We consider two specific choices of factor loadings:

Case I: The factor loadings are taken to be $a_j = 0.25$ and $b_j = 0.1$ for all j. Thus, ϵ_{ij}'s have the same marginal distribution, although they are correlated.

Case II: The factor loadings $\{a_i\}$ and $\{b_i\}$ are generated independently from, respectively, $U(0, 0.4)$ and $U(0, 0.2)$.

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:

$$c\,\delta_0 + \tfrac{1}{2}\,(1-c)\,\exp(-|x|)\,,$$

where $c \in (0, 1)$ is a constant. With the noise and the expression level given above, $\{Y_{ij}\}$ generated from (2.1) 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 given α_ν, we compute the P-value according to the normal approximation, t-approximation, the bootstrap method and the aggregated bootstrap (4.2). This results in ν estimated P-values $\{\hat{p}_j\}$ for each method and each simulation. Let N denote the number of P-values that are no larger than α_ν; see (1.3). Then, N/ν is the empirical fraction of the null hypotheses that are rejected. When $c = 0$, $N/(\nu\alpha_\nu) - 1$ reflects the accuracy of approximating the P-values, and its the root mean square error (RMSE), $\{E(N/(\nu\alpha_\nu) - 1)^2\}^{1/2}$, will be reported, where the expectations are approximated by the averages across simulations. We exclude the marginal aggregation method (4.1), since ir out simulations it always gave $N/\nu = \alpha_\nu$.

We take $\nu = 600$ (small), $\nu = 1,800$ (moderate) and $\nu = 6,000$ (typical) for microarray applications (after preprocessing, which filters out many low quality measurements on certain genes) and $\alpha_\nu = 1.5\nu^{-2/3}$, resulting in $\alpha_\nu = 0.02$, 0.01 and

0.005, respectively. The sample size n is taken to be 6 (typical number of microarrays), 20 (moderate) and 50 (large), and the number of replications in simulations is 600,000/ν. For the bootstrap calibration method and the aggregated method (4.2), we replicate bootstrap samples 2,000, 4,000 and 9,000 times, respectively, for $\alpha_\nu = 0.02, 0.01$ and 0.005.

Tables 1 and 2 report the accuracy of estimated P-values when $c = 0$. In reading these tables, keep in mind the computation involved in simulating estimated tail probabilities for the bootstrap method. For example, using the bootstrap method, we need to compute $600{,}000 \times 9{,}000 = 5.4 \times 10^9$ t-statistics of sample size n, yet in computing RMSE $\{E(N/(\nu\alpha_\nu) - 1)^2\}^{1/2}$ at $\nu = 6{,}000$, we merely calculate the expectation $\{E(N/30 - 1)^2\}^{1/2}$ ($\nu\alpha_\nu = 30$) over 100 simulations.

Tables 1 and 2 are about here

First of all, the normal approximations are too inaccurate to be useful. Therefore we shall exclude the normal method in the discussion below. For $n = 20$ and 50, the bootstrap method provides better approximations than Student's t-method. This indicates that the bootstrap can test more hypotheses simultaneously, which is in accord with our asymptotic theory on the accuracy of approximations of P-values. Overall the bootstrap method is also slightly better the aggregated bootstrap (4.2), although the two methods are effectively comparable. However with the small sample size $n = 6$, Student's t-method is relatively the best, although the approximations are poor in general. This is understandable, as the noise distribution is not normal. With such a small sample size, the two bootstrap based methods, in particular the aggregated bootstrap method (4.2), suffer more from the random fluctuation in the original samples.

To examine the FDR, we repeat the above experiment but with $c = 0.5$. Since the normal approximation is too inaccurate, we compare only the three other methods, i.e. those based on the t-distribution, the bootstrap and bootstrap aggregation. We set the control level at $p = 0.25$ in (1.2). The number of falsely discovered genes

is given by

$$N = \sum_{i=1}^{n} I(p_i \le p_{(k_n)}, \mu_i = 0)\,.$$

The boxplots of the FDR, N/k_n, are presented in Fig. 1 with $a_i = 0.25$ and $b_i = 0.1$, and in Fig. 2 with a_i independently drawn from $U(0,\, 0.4)$ and b_i independently drawn from $U(0,\, 0.2)$, where k_n is defined as in (1.2).

With sample size $n = 20$ and 50, the actual FDR is far below the control level $p = 0.25$. This is also true for $n = 6$ with both t and bootstrap methods. This may be explained as follows. According to Benjamini and Hochberg (1995), the expected value of FDR is bounded from above by $\nu_0 p/\nu$, where ν_0 is the number of true null hypotheses. In our setting, $\nu_0/\nu \approx 0.5$ (as $c = 0.5$). Hence, the expected FDR should be bounded by 12.5%. This is indeed the case for the bootstrap calibration methods when $n = 20$ and 50. It is also worthwhile noting that with $n = 20$ and 50 the FDRs obtained from the bootstrap and aggregated bootstrap tend to be smaller than those obtained from the t-distribution. Also the random fluctuation is severe when sample size is $n = 6$. This may explain why the FDR based on aggregated bootstrap (4.2) is highly volatile.

Finally we conduct a separate study to compare between k'_n defined in (1.3), and $\hat{k}_n$ defined in (1.4). We draw 100 samples from the model used in Fig. 1 above, with different values of n and ν. Table 3 lists the relative frequencies of $\hat{k}_n - k'_n$ taking values $0, 1, 2, \ldots$ in a simulation with 100 replications. As one would expect, $\hat{k}_n$ and k'_n differ from each other with very small probabilities, especially when the sample size n is 20 and 50. In particular, the Benjamini and Hochberg method, with empirical FDR, rarely discovered more than 2 genes than the traditional method (1.3).

6. PROOFS OF RESULTS IN SECTIONS 3 AND 4

6.1. Auxiliary result.

Let $C_1 > 0$. Given a random variable X with $E(X) = 0$, consider the condition:

$$E(X) = 0\,, \quad E\big(X^2\big) = 1\,, \quad E\big(X^4\big) \le C_1\,. \tag{6.1}$$

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

Theorem 6.1. *Let $X, X_1, X_2, \ldots$ denote independent and identically distributed random variables such that (6.1) holds. Write $T = T(n)$ for Student's t statistic computed from the sample $X_1, \ldots, X_n$, 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 $X/(\mathrm{var}X)^{1/2}$. Then,*

$$\frac{P(T > x)}{1 - \Phi(x)} = \exp\left(\pi_3\, x^3\, n^{-1/2}\right) \left\{1 + \theta\, \frac{(1+x)^2}{n^{1/2}}\right\}, \tag{6.2}$$

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

6.2. Proof of Theorem 3.1.

Theorem 3.1 in the case of normal calibration follows directly from Theorem 6.1. The case of Student's t calibration can be treated similarly. In either setting, to conduct the calculations leading to (3.2) note that, if $x_n \sim cn^{1/6}$ and the random variable N is distributed as normal $\mathrm{N}(0, 1)$, then we have uniformly in i,

$$\frac{P(|T_i| > x_n)}{P(|N| > x_n)} = \cosh\left(\tfrac{1}{3}\,\kappa_{i3}\, c^3\right) + o(1)\,.$$

In the case of normal calibration we choose $x_n = t_\alpha$ to solve the first equation in (2.3), implying that $P(|N| > t_\alpha) \sim -\nu^{-1}\, \log(1-\alpha)$. Hence, $t_\alpha \sim (2 \log \nu)^{1/2}$, so that the condition $x_n \sim cn^{1/6}$ is equivalent to $n^{-1/3} \log \nu \to \frac{1}{2} c^2$.

6.3. Proof of Theorem 3.3.

The basic idea is to show that T^* has expansion (6.2) for samples $\mathcal{Y}_i$ falling in a set $\mathcal{E}_n$, the probability of which tends to 1 at an exponential rate. This follows by checking the conditions in Theorem 6.1 for bootstrap samples for all $\mathcal{Y}_i \in \mathcal{E}_n$.

Note that each $\mathrm{var}(\epsilon_i') = 1$. To check that, with probability at least $p_n \equiv 1 - \exp(-d_1\, n^{1/2})$ for a constant $d_1 > 0$, the conditions of Theorem 6.1 hold for the bootstrap distribution of the statistic T_i^*, for each $1 \le i \le \nu$, 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 $1 \le i \le \nu$:

$$C_4 \le \frac{1}{n} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2\,, \quad \frac{1}{n} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^4 \le C_5\,. \tag{6.3}$$

This can be done using Bernstein's inequality and the assumption that, for each i, $P(|\epsilon_i'| \le C) = 1$, and can also be shown by the uniform convergence result of the empirical process of Korosok and Ma (2005).

Let $\mathcal{E}_n$ denote the event that (6.3) holds for each $1 \le i \le \nu$. When $\mathcal{E}_n$ prevails, we may apply Theorem 6.1 to the distribution of T_i^* conditional on $\mathcal{Y}_i$, obtaining:

$$P\big(T_i^* > x \;\big|\; \mathcal{Y}_i\big) = \{1 - \Phi(x)\} \exp\big(-\tfrac{1}{3}\,\hat{\kappa}_{i3}\, n^{-1/2}\, x^3\big) \left\{1 + \hat{\theta}_i\, \frac{(1+x)^2}{n^{1/2}}\right\}, \tag{6.4}$$

where $\hat{\kappa}_{i3}$ is the empirical version of κ_{i3}, computed from $\mathcal{Y}_i$, and, on an event of which the probability equals $1 - O\{\exp(-d_2\, n^{1/2})\}$, $|\hat{\theta}_i| \le D_1$ uniformly in i and in $0 \le x \le x_n$. (Here and below, x_n will denote any sequence diverging to infinity but satisfying $x_n = o(n^{1/4})$, and $D_1, D_2, \ldots$ and $d_1, d_2, \ldots$ will denote constants.) It follows directly from Theorem 6.1 that

$$P_{0i}(T_i > x) = \{1 - \Phi(x)\} \exp\big(-\tfrac{1}{3}\,\kappa_{i3}\, n^{-1/2}\, x^3\big) \left\{1 + \theta\, \frac{(1+x)^2}{n^{1/2}}\right\}, \tag{6.5}$$

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

Result (6.5), and its analogue for the left-hand tail of the distribution of T_i, allow us to express $t_{i\alpha}$, the solution of the equation $P_{0i}(|T_i| > t_{i\alpha}) = 1 - (1-\alpha)^{1/\nu}$, as a Taylor expansion:

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

uniformly in i, where c is a constant and z_α is the solution of $P(|N| > z_\alpha) = 1 - (1-\alpha)^{1/\nu}$. Note that if z_α solves this equation then $z_\alpha \sim (2 \log \nu)^{1/2}$, and

so, since $\log \nu = o(n^{1/2})$, then $z_\alpha = o(n^{1/4})$. Therefore, without loss of generality, $0 \le z_\alpha \le x_n$. Likewise we may assume below that $0 \le t_{i\alpha} \le x_n$, and $0 \le \hat{t}_{i\alpha} \le x_n$ with probability $1 - O\{\exp(-d_2\, n^{1/2})\}$.

Also, from (6.4) we can see that on an event of which the probability equals $1 - O\{\exp(-d_2\, n^{1/2})\}$,

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

However, on an event with probability $1 - O\{\exp(-d_3\, n^{1/2})\}$, $|\hat{\kappa}_{i3} - \kappa_{i3}| \le D_4\, n^{-1/4}$, and therefore, on an event with probability $1 - O\{\exp(-d_4\, n^{1/2})\}$,

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

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

$$P_{0i}(|T_i| > t_{i\alpha} \pm \delta) \mp D_6\, \exp\left(-d_4\, n^{1/2}\right), \tag{6.6}$$

where

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

Using (6.5), and its analogue for the left-hand tail, to expand the probability in (6.6), we 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 $D_7 > 0$. Hence,

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

uniformly in i. Theorem 3.3 follows from this property.

6.4. Proofs of Theorems 4.1 and 4.2.

The proofs of Theorems 4.1 and 4.2 are similar. Since Theorem 4.1 is more involved, we outline here the proof of that result.

Note that

$$E\hat{F}_\nu(x_n) = \nu^{-1} \sum_{i=1}^{\nu} P(T_i \le x_n) = F(x_n) + O(\nu_1/\nu)\,.$$

Let $\mathcal{I}$ be the set of indices $\mathcal{I}$ for which H_{0i} is true. Note that if $i \in \mathcal{I}$ then, by the assumption of identical null distribution, $\mathrm{var}\{I(T_i \le x_n)\} = F(x_n)\,\{1 - F(x_n)\}$, and for $i \notin \mathcal{I}$, the variance is bounded above by $1/4$. Using these results we have:

$$\mathrm{var}\{\hat{F}_n(x_n)\} = \nu^{-1}\,F(x_n) + \nu^{-2} O\Big\{\nu_1 + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}} F(x_n)\,|r_{ij}| + \sum_{i\in\mathcal{I}}\sum_{j\notin\mathcal{I}} F(x_n)^{1/2}\,|r_{ij}| + \nu_1^2\Big\}\,.$$

The second term is bounded by

$$O\big\{F(x_n)\,a_n + F(x_n)^{1/2}\,\nu_1\,a_n\,\nu^{-1} + \nu_1^2\,\nu^{-2}\big\}\,.$$

The result now follows from a standard mean-variance decomposition.

REFERENCES

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 YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–1188.

BENTKUS, V. AND GÖTZE, F. (1996). The Berry-Esseen bound for Student's statistic. *Ann. Statist.* **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.

DUDOIT, S., SHAFFER, J.P. AND BOLDRICK, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18**, 71–103.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99**, 96–104.

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

FAN, J. and LI, R. (2006). Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery Proceedings of the Madrid International Congress of Mathematicians 2006, to apepar.

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

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

GENOVESE, C. AND WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035–1061.

He, X. and Wang, H. (2006). Detecting differential expressions in GeneChip microarray studies: A quantile approach. *Manuscript.*

HU, J. and HE, X. (2006). Enhanced quantile normalization of microarray data to reduce loss of information in the gene expression profile. *Biometrics*, to appear.

HUANG, J., WANG, D., AND ZHANG, C. (2005). A two-way semi-linear model for normalization and significant analysis of cDNA microarray data. *J. Amer.*

Statist. Assoc. **100**, 814–829.

HUBER, P.J. (1973). Robust regression : asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, **1**, 799–821.

KOROSOK, M.R. AND MA, S. (2005). Marginal asymptotics for the "large p, small n" paradigm: With applications to micorarray data. *Manuscript*.

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

LEHMANN, E.L. AND ROMANO, J.P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33**, 1138–1154.

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

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, **16**, 356-366.

REINER, A., YEKUTIELI, D. AND BENJAMINI, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368-375.

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

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. *Elect. J. Probab.* **10**, 1260–1285.

Table 1: Root mean squared errors of $N/(\nu\alpha) - 1$. In model (5.1), $a_i \equiv 0.25$ and $b_i \equiv 0.1$.

	$n = 6$			$n = 20$			$n = 50$		
α	0.02	0.01	0.005	0.02	0.01	0.005	0.02	0.01	0.005
Normal	3.425	5.604	9.083	0.833	1.221	1.768	0.388	0.528	0.696
t	0.459	0.494	0.512	0.258	0.329	0.391	0.242	0.284	0.313
Bootstrap	0.546	0.644	0.657	0.201	0.282	0.296	0.224	0.250	0.244
(4.2)	0.842	0.946	0.990	0.202	0.297	0.352	0.228	0.249	0.262

Table 2: Root mean squared errors of $N/(\nu\alpha) - 1$. In model (5.1), $a_i \sim U(0, 0.4)$ and $b_i \sim U(0, 0.2)$.

	$n = 6$			$n = 20$			$n = 50$		
α	0.02	0.01	0.005	0.02	0.01	0.005	0.02	0.01	0.005
Normal	3.351	5.596	9.014	0.770	1.189	1.707	0.339	0.526	0.526
t	0.406	0.485	0.456	0.307	0.273	0.347	0.182	0.299	0.299
Bootstrap	0.564	0.637	0.677	0.202	0.262	0.322	0.162	0.284	0.284
(4.2)	0.851	0.941	0.985	0.201	0.289	0.379	0.165	0.278	0.278

Table 3: Relative frequencies of $\widehat{k}_n - k_n^*$ taking different values.

	$\nu = 600,\ \alpha = 0.02$			$\nu = 1800,\ \alpha = 0.01$			$\nu = 6000,\ \alpha = 0.005$		
$\widehat{k}_n - k_n^*$	$n = 6$	$n = 20$	$n = 50$	$n = 6$	$n = 20$	$n = 50$	$n = 6$	$n = 20$	$n = 50$
0	0.73	0.85	0.89	0.75	0.85	0.94	0.69	0.91	0.96
1	0.20	0.08	0.09	0.13	0.12	0.06	0.16	0.07	0.03
2	0.02	0.06	0.01	0.07	0.02	0.00	0.07	0.02	0.01
3	0.04	0.01	0.01	0.02	0.01	0.00	0.02	0.00	0.00
≥ 4	0.01	0.00	0.00	0.03	0.00	0.00	0.06	0.00	0.00

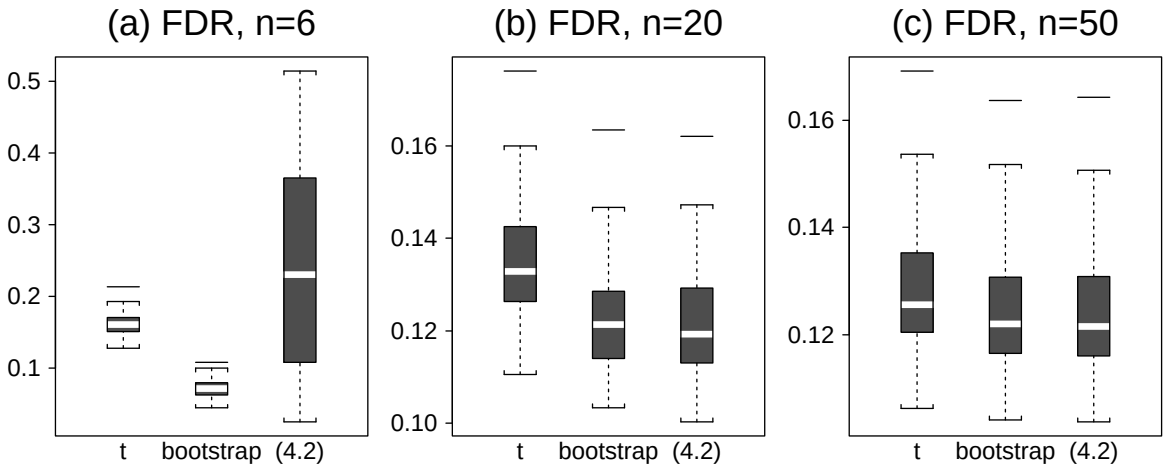

Figure 1: Boxplots of FDR obtained based on t-distribution (t), bootstrap method (bootstrap), and marginal aggregation (4.2). In model (5.1), $a_i \equiv 0.25$ and $b_i \equiv 0.1$.

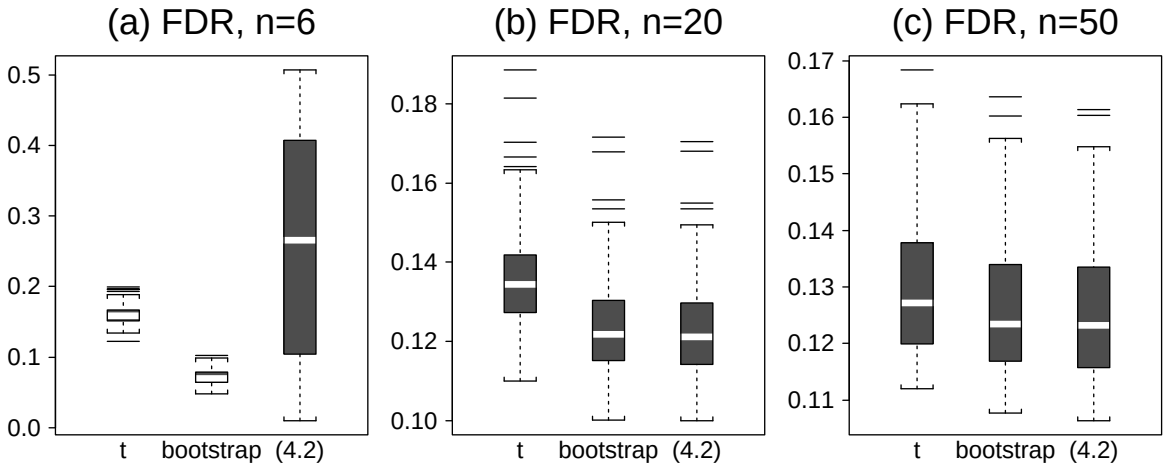

Figure 2: Boxplots of FDR obtained based on t-distribution (t), bootstrap method (bootstrap), and marginal aggregation (4.2). In model (5.1), $a_i \sim U(0, 0.4)$ and $b_i \sim U(0, 0.2)$.