Journal of Statistical Research *ISSN 0256 - 422 X*
2010, Vol. 44, No. 1, pp. 85-102
Bangladesh

CONTROLLING THE AVERAGE FALSE DISCOVERY IN LARGE-SCALE MULTIPLE TESTING

Muni S. Srivastava
Department of Statistics, University of Toronto, Toronto, ON, Canada
Email: srivasta@utstat.toronto.edu

SUMMARY

In this paper, we consider multiple testing procedures in which we simultaneously test a large number m of null hypotheses $H_1, \ldots, H_m$ using the test statistics $T_1, \ldots, T_m$. The currently used procedure of controlling the false discovery rate (FDR) at a specified level requires that the statistics $T_1, \ldots, T_m$ be either independently distributed or positively related. In practice T_i's are rarely independent and it is not known how to ascertain the positive relationship between T_i's. In this paper, we propose to control the expected value of the Average False Discovery (AFD) at some specified level. This AFD procedure controls its level at the specified value independent of how T_i's are related. This specified value can be chosen to control k-FWER or γFWER and even FDR at their respective specified levels. Using simulation, we compare our proposed AFD procedure with the FDR procedure. In terms of power and stability, the proposed AFD procedure has an edge over the FDR procedure, as well as over k-FWER procedure. Two illustrative examples are given to compare the number of differentially expressed genes obtained by the two methods.

Keywords and phrases: Familywise error rate, false discovery rate, microarray datasets, multiple testing, power, sample size smaller than dimension, variability.

1 Introduction

One of the most important issue in large-scale multiple testing is how to choose or even define the type I error rate, usually called level of significance, although no such problem arises if our interest is only in testing globally the hypothesis $H = \bigcap_{i=1}^{m} H_i$ against the alternative that at least one H_i is false. This alternative will be written $A \neq H$. In this case, the type I error rate is well defined and is given by

$$\mathrm{P}(\text{rejecting } H | H \text{ is true}) \leq \alpha \ ,$$

where α is a pre-assigned number, $0 < \alpha < (1/2)$. This is also called Familywise Error Rate or simply FWER. However, when the hypothesis H is rejected, we need to know, which H_i or H_i's may have caused the rejection of the hypothesis H. For small m, we

may choose α/m as a significance level for each individual hypothesis and maintain, albeit conservatively, the significance level α by using the Bonferroni inequality, which does not require the independence of the statistics T_i's used in testing the hypothesis H_i, But when m is large, α/m becomes too small. Even if T_i's are independently distributed, the exact level γ/m at which each hypothesis H_i may be tested is obtained by solving the equation

$$1 - (1 - \gamma/m)^m = \alpha.$$

For large m, $\gamma \simeq \alpha$. Thus, it is not the conservativeness of the Bonferroni inequality that causes the problem, but rather the largeness of m. Thus, we shall consider other kinds of 'Type I' errors which may not necessarily translate into finding confidence intervals for each parameter at a pre-assigned confidence level. To define other kinds of type I errors in connection with simultaneously testing m hypotheses H_i'a using the test statistics T_i's, let m_0 be the number of true null hypotheses. Then, there are $m_1 = m - m_0$ false hypotheses, but both the numbers m_0 and m_1 are unknown. In fact, it is not known which m_0 of the H_i's are true hypotheses. Similarly, we do not know which m_1 of the m hypotheses are false hypotheses. We shall denote by Λ_0, the set of m_0 true null hypotheses, and by Λ_1 the set of m_1 false hypotheses. Let R be the number of total rejections which contain a total of V false rejections (of the true hypotheses), and $R - V$ correct rejections (of the false hypotheses). But V is not observable and we only observe R and m. It is customary to represent these numbers as in the following table.

Number of	Rejected	Not Rejected	Total
True null	V	$m_0 - V$	m_0
False null	S	$m_1 - S$	m_1
Total	R	$m - R$	m

Table 1: The outcomes of a multiple testing procedure.

We first note that the familywise error rate FWER mentioned earlier is defined by

$$FWER = P_{\Lambda_0}(V \geq 1). \tag{1.1}$$

which is controlled at a specified level of significance, say α, usually achieved by Bonferroni inequality. Benjamini and Hochberg (1995) introduced and defined a type I error, called False Discovery Rate, FDR. It is given by

$$\begin{aligned} FDR &= \mathrm{E}[(V/R)I(R > 0)] \\ &= \mathrm{E}[(V/R)|R > 0]P\ (R > 0), \end{aligned} \tag{1.2}$$

where $I(\cdot)$ denotes the indicator function in which $I(R > 0) = 1$ if $R > 0$ and takes the value zero if $R = 0$; it may be noted that when $m_0 = m, FDR = FWER$. The expected value in (1.2) is obtained when true null holds for V and false null holds for $S = R - V$.

Benjamini and Hochberg showed that the FDR can be controlled at any specified level, say, α by using the step-up procedure of Simes (1986), provided the statistics T_i's used for testing the hypotheses H_i are independently distributed. Later, Benjamini and Yekutieli (2001) showed that the FDR can be controlled if T_i's are independently distributed or positively related. And if these conditions are not satisfied, then FDR is controlled at level $\alpha \sum_{j=1}^{m} (1/j)$. Thus, in the general case the boundaries of Sime's procedure have to be modified to control FDR at level α, making FDR less powerful To distinguish it from other procedures, we shall call the above procedure due to Benjamini and Hochberg (1995) as the BH-FDR procedure.

Hommel and Hoffman (1988), and Lehmann and Romano (2005) considered to control another kind of Type I error by proposing to control,

$$P_{\Lambda_0} \{V \geq k\} \leq \alpha \tag{1.3}$$

for some chosen k. This can be controlled by using constant rejection region, no matter how the statistics T_i's are related, as opposed to the varying rejection regions of the FDR. Romano and Shaikh (2006) extended this result to varying critical regions, similar to FDR, which improved the power but the level of α is rarely achieved for both the cases. But, the problem of how to choose k in (1.3) remains. To overcome this difficulty Du and Srivastava (2006) proposed to control

$$P_{\Lambda_0} \left\{ \frac{V}{m_0} \geq \gamma \right\} \leq \alpha \tag{1.4}$$

and called it γFWER procedure, and chose $\gamma = .05$, and 0.10, α=0.05 and compared it with the BH-FDR procedure at level $\alpha = 0.05$. For $(m_0/m) \geq 0.7$, the procedure (1.4) has a better power than the BH-FDR procedure while the FDR of the (1.4) procedure remained at or about the same level as BH-FDR.

The procedures given in (1.2) - (1.4) do not control the number of false discoveries V. While it may be difficult to control it in large scale hypotheses testing, it may be possible to control the expected value of V/m_0, since neither V nor m_0 is observed, almost similar to the FDR case which is also the expected value of an unobserved random variable. Thus, in this paper, we propose to control

$$AFD = \mathrm{E}_{\Lambda_0}(V/m_0) = \frac{1}{m_0} \sum_{j \in \Lambda_0} \mathrm{P}(reject\ H_j | H_j). \tag{1.5}$$

which we call Average False Discovery procedure. It will be shown in Section 2, that the AFD can be controlled by a fixed rejection region procedure at any specified level, say, δ, irrespective of how the statistics T_i's used for testing the hypothesis H_i are related.

It may be noted that when $m_0 = m$, the AFD defined above becomes PCER, the per comparison error rate, see, Dudoit et al.(2003), for example.

It is shown in Section 2.4, that the control level of the AFD procedure can be chosen to control the level of the k-FWER procedure also, but it can not be chosen to control the level of PCER procedure unless $m_0 = m$.

Next, we discuss the selection of the values of δ required to carry out the AFD procedure. Clearly, it is no different than the selection of 'α' in FWER or FDR procedures. In Section 2.3, we give a method to choose δ so that the FDR can also be controlled at some specified level when $T_i's$ are independent and m is large. In fact, in all our simulations, the FDR of the proposed AFD procedure is controlled even when $T_i's$ are not independent. In our simulation, we choose $\delta = 0.005$ in order that the FDR be controlled at level 0.05. Thus, the proposed procedure can control both AFD and FDR by choosing δ appropriately as given in Section 2.3 . Alternatively, δ may be chosen to control k-FWER or γFWER as shown in Section 2.4. In any case, δ may be chosen so that there are no more than 5 or 10 per thousand of false discoveries.

The organization of this article is as follows. In Section 2, we describe the proposed procedure along with some properties of this procedure. In Section 3, we describe the BH-FDR procedure of Benjamini and Hochberg (1995). The power of the two procedures is compared in Section 4 and two examples of microarrays are analyzed in Section 5. The paper concludes in Section 6.

2 The Average False Discovery and Other Related Procedures.

Genovese and Wasserman (2002) showed that asymptotically for large m, the BH-FDR procedure in the independent statistics case, is equivalent to a constant critical region procedure obtained at some level ε, $0 < \alpha/m \leq \varepsilon \leq \alpha < \frac{1}{2}$ but it is not known how to choose ε. The proposed AFD procedure thus fills this gap.

To motivate the AFD procedure, let us recall that in a single testing situation, one tries to control the probability of rejecting the hypothesis H when it is actually true. Let $V = 1$ if the hypothesis H is rejected and $V = 0$ otherwise. Then the type I error rate in a single hypothesis testing situation can be written as

$$\mathrm{P}(\textit{reject } H|H) = \mathrm{P}(V = 1|H) = \mathrm{E}[V|H]. \tag{2.1}$$

Suppose T is an appropriate test statistic for H. Suppose t is an observed value of T and large values of t are evidence against H. Then the p-value of T at t is given by

$$p = \mathrm{P}(T \geq t|H). \tag{2.2}$$

Consider the procedure that rejects H if $p \leq \alpha$ for a specified value α. Then this testing procedure has a significance level α. In other words, the type I error rate is controlled at level α.

In the multiple testing situation, let us consider the average of the type I error rates of all of the true cases. That is,

$$\frac{1}{m_0}\mathrm{E}_{\Lambda_0}[V] = \frac{1}{m_0}\sum_{j\in\Lambda_0}\mathrm{P}(reject\ H_j|H_j), \tag{2.3}$$

where Λ_0 is the set of the true null hypotheses.

Suppose that $T_1,\ldots,T_m$ are appropriate test statistics for the multiple hypotheses $H_1,\ldots,H_m$, and $t_1,\ldots,t_m$ their observed values. Large values of t_j's are evidence against H_j's. Then the p-values of $T_1,\ldots,T_m$ are given by

$$p_j = \mathrm{P}(T_j \geq t_j|H_j). \tag{2.4}$$

When the p-values are considered as random variables, we denote them by P_j's. It will be assumed that for any $0 < \nu \leq 1$,

$$\mathrm{P}(P_j \leq \nu|H_j) = \nu, \tag{2.5}$$

which holds, when the samples are taken from the continuous model.

Consider the procedure of rejecting the hypothesis H_j if $p_j \leq \delta$, where δ is a pre-specified value. Then

$$\begin{aligned} AFD &= \frac{1}{m_0}\mathrm{E}[V|\Lambda_0] = \frac{1}{m_0}\mathrm{E}\left[\sum_{j\in\Lambda_0} I(P_j \leq \delta)|\Lambda_0\right] \\ &= \frac{1}{m_0}\sum_{j\in\Lambda_0}\mathrm{P}(P_j \leq \delta|H_j) = \frac{1}{m_0}\sum_{j\in\Lambda_0}\delta = \delta, \end{aligned} \tag{2.6}$$

where $I(\cdot)$ is the indicator function. Therefore, a procedure that rejects H_j when $p_j \leq \delta$ controls the AFD at level δ. Values of δ can be chosen in the same manner as the value of α; it does not depend on the value of m, m_0 or m_1.

Theorem 2.1. A procedure that rejects each hypothesis H_j if the corresponding p-value $p_j \leq \delta$ controls the average false discovery at level δ. That is

$$AFD = \delta. \tag{2.7}$$

It may be noted that $E[V|\Lambda_0] = m_0\delta \leq m\delta$, which provides a good bound for the average of false discoveries as in most practical situations $m_0/m \geq 0.90$.

In the next four subsections, we study the properties of the AFD procedure. We start with the power.

2.1 Power of the AFD procedure

We consider the average power

$$\pi = \mathrm{E}[S]/m_1, \tag{2.8}$$

where S is the number of false hypotheses that have been rejected and m_1 is the total number of false hypotheses.

Without loss of generality, suppose that the first $H_1, \ldots, H_{m_1}$ are false hypotheses. And assume that the alternative hypotheses $A_1, \ldots, A_{m_1}$ are equal, that is,

$$X_1, \ldots, X_{m_1} \sim A.$$

Then, when X_j's are continuous random variables, the average power is given by

$$\begin{aligned} \pi = \mathrm{E}[S/m_1] &= \frac{1}{m_1} \sum_{j \notin \Lambda_0} \mathrm{E}[I(P_j \le \delta) | A_j] \\ &= \frac{1}{m_1} \sum_{j \notin \Lambda_0} \mathrm{P}(P_j \le \delta | A) = \frac{1}{m_1} \sum_{j \notin \Lambda_0} F(\delta) = F(\delta), \end{aligned} \tag{2.9}$$

where $F(\delta) = \mathrm{P}(P_j \le \delta | A)$ is unknown, but depends on the alternative distribution of $X_1, \ldots, X_{m_1}$. This implies that the average power of the proposed AFD procedure stays constant irrespective of the number m_1 when the alternative distributions are the same. This result is of great significance since some analysts may choose only a smaller number $m^* \le m$ of important hypotheses, which obviously may have smaller number $m_1^* \le m_1$ of false hypotheses. Then, it is important that the power should not be dependent on m_1.

The above result is summarized in the following theorem.

Theorem 2.2. Assume that all the m_1 alternative hypotheses have the same distribution. Then the power of the AFD procedure remains constant.

2.2 Variance of V/m_0

The variance of V/m_0 is given by the following two lemmas, with the proofs given in the Appendix.

Lemma 2.1. When the samples are taken from continuous models, under general conditions,

$$\mathrm{Var}_{\Lambda_0}(V/m_0) = \frac{\delta}{m_0} + \frac{1}{m_0^2} \sum_{j,l \in \Lambda_0, j \ne l} \mathrm{E}_{\Lambda_0}[I(P_j \le \delta) I(P_l \le \delta)] - \delta^2.$$

Lemma 2.2. Let the test statistics are student's t-statistics with the distribution function G. Let $G^{-1}(1-\delta) = t_0$. Then for large N,

$$\mathrm{Var}_{\Lambda_0}(V/m_0) \simeq \frac{\delta}{m_0} + \frac{1}{m_0^2} \sum_{j,l \in \Lambda_0, j \ne l} \psi(\delta, \rho_{jl}) - \delta^2,$$

where ρ_{jl}'s are the correlation coefficients between the test statistics, and

$$\psi(\delta, \rho_{jl}) = \frac{1}{\sqrt{2\pi}} \int_{|x| \ge t_0} \left[\Phi\left(\frac{-t_0 + \rho_{jl} x}{\sqrt{1-\rho_{jl}^2}} \right) + \Phi\left(\frac{-t_0 - \rho_{jl} x}{\sqrt{1-\rho_{jl}^2}} \right) \right] \exp(-x^2/2) dx,$$

with $\Phi(\cdot)$ being the standard normal distribution function.

From above, we get the following corollary.

Corollary 2.1. When the test statistics are independently distributed,

$$\mathrm{Var}_{\Lambda_0}(V/m_0) = \delta(1-\delta)/m_0.$$

2.3 Controlling the FDR of the AFD Procedure under the Independence of T_i's or P_i's

In this section, we shall assume that the statistics T_i's for testing the hypotheses H_i's are independently distributed. Thus the p-values are independently distributed. Suppose we wish to control the FDR of the BH procedure at level α. We shall assume that

$$0 < \lim_{m\to\infty} (m_i/m) < 1,\ i = 0, 1.$$

Then from the law of large numbers

$$\frac{V}{m_0} = \frac{1}{m_0}\sum_{i\in\Lambda_0} I(P_i < \delta) \xrightarrow{p} \mathrm{P}_{\Lambda_0}(P_i \le \delta) = \delta.$$

Similarly,

$$\frac{S}{m_1} = \frac{1}{m_1}\sum_{i\in\Lambda_1} I(P_i < \delta) \xrightarrow{p} \mathrm{P}_{\Lambda_1}(P_i \le \delta) = F(\delta),$$

under the assumption that when $i \in \Lambda_1$ all the P_i are independently and identically distributed with common cumulative distribution function F. Also, since the random variable $\frac{V}{R}$ is a decreasing function of R and the random variable $I(R > 0)$ is an increasing function of R, it follows from Theorem 1.10.5 of Srivastava and Khatri (1979, page 26) that

$$FDR = E\left[\left(\frac{V}{R}\right) I(R > 0)\right] \le E\left(\frac{V}{R}\right) P(R > 0)) \le E\left(\frac{V}{R}\right)\ ,$$

where $R = V + S$. Hence,

$$\begin{aligned} FDR \le E\left(\frac{V}{R}\right) &= \mathrm{E}\left[\frac{\sum_{i\in\Lambda_0} I(P_i \le \delta)}{\sum_{i\in\Lambda_0} I(P_i \le \delta) + \sum_{i\notin\Lambda_0} I(P_i \le \delta)}\right] \\ &= \frac{\delta}{\delta + \frac{m_1}{m_0}F(\delta)} + O(m^{-\frac{1}{2}})\ . \end{aligned}$$

Thus, the FDR can be controlled asymptotically at level α if

$$\frac{\delta}{\delta + \frac{m_1}{m_0}F(\delta)} \le \alpha\ ,$$

or equivalently if

$$(1-\alpha)\delta \le \alpha\frac{m_1}{m_0}F(\delta)\ . \tag{2.10}$$

That is, if

$$F(\delta) \geq \frac{m_0}{m_1}\frac{1-\alpha}{\alpha}\delta. \tag{2.11}$$

In most practical situations, $m_1/m \leq 0.20$. Since m_1, would never be known, it would be prudent to guard against all these situations and choose $(m_0/m_1) = 4$. Thus, we shall chose δ so that $F(\delta) \geq 4\frac{(1-\alpha)}{\alpha}\delta$. For $\alpha = 0.05$ and $\delta = 0.005$, $F(0.005) \geq 0.38$, which is satisfied in all our simulations. Thus, for large m, FDR is controlled at level 0.05, when AFD is controlled at level 0.005 for the AFD procedure. If a good knowledge of m_1 is available, other choices of δ can also be made. The above suggestion is on the conservative side.

2.4 Choosing δ to control k-FWER and γFWER.

From the Markov inequality see, e.g. Lehmann and Romano (2005), we have

$$P_{\Lambda_0}\{V > k\} \leq \frac{E_{\Lambda_0}(V)}{k} = \frac{m_0}{k}AFD \leq \frac{m}{k}AFD\ .$$

Thus to control k-FWER at level α, we require that

$$AFD \leq \frac{k}{m}\alpha = \gamma\alpha$$

where $\gamma = k/m$. Thus, in order to control k-FWER or γFWER at level α, we need to choose δ less than $(k/m)\alpha = \gamma\alpha$. This also shows that the k-FWER procedure and its variants are controlled conservatively at level α, while AFD procedure is controlled exactly at level α.

3 Benjamini-Hochberg Procedure Controlling FDR

Benjamini and Hochberg (1995) introduced the concept of False Discovery Rate, FDR. Define a random variable η by

$$\eta = (V/R)I(R > 0). \tag{3.1}$$

Then as in (1.2), the FDR is given by

$$FDR = \mathrm{E}[\eta]. \tag{3.2}$$

It may be noted that since V is unobservable, η is also an unobservable random variable. However, Benjamini and Hochberg (1995) showed that the FDR can be controlled at a specified level, say, α by using Simes (1986) step-up procedure described as follows, provided that the statistics T_i's used in testing the hypotheses H_i's are independently distributed. Benjamini and Yekutieli (2001), however, have relaxed the independence condition to positive relationship between T_i's. Let

$$p_i = \mathrm{P}\{T_i > t_i | H_i\} \tag{3.3}$$

be the observed p-value for the observaed value of the statistics t_i. Let

$$p_{(1)} \leq \cdots \leq p_{(m)} \tag{3.4}$$

be the ordered values of the p_i's. As before, when p_i and $p_{(i)}$ are random, they will be denoted by P_i and $P_{(i)}$ respectively. The hypothesis corresponding to $P_{(i)}$ will be denoted by $H_{(i)}$. Then according to Simes' procedure, the hypotheses $H_{(1)}, \ldots, H_{(j^*)}$ are rejected if j^* is the largest value of j for which $p_{(j)} \leq j\alpha/m$. That is

$$j^* = max_{1 \leq j \leq m}\{j : p_{(j)} \leq j\alpha/m\} \ . \tag{3.5}$$

If $T_1, \ldots, T_m$ are independently distributed or positively related, then

$$FDR = \mathrm{E}_\Lambda(\eta) \leq (m_0/m)\alpha \leq \alpha. \tag{3.6}$$

However, if T_i's are not independently distributed or positively related, then j^* is defined by

$$j^* = \max_{1 \leq j \leq m} \left[j : p_{(j)} \leq \frac{j\alpha}{mC_m} \right] \ , \ C_m = \sum_{i=1}^{m} \left(\frac{1}{i} \right) \tag{3.7}$$

for (3.6) to hold. From (3.6), it follows that for large m_1 and m, it is a very conservative procedure. Thus, it will have less power for large m_1 and m. The power of any procedure is defined by

$$\mathrm{E}_{\Lambda_1} \left(\frac{S}{m_1} \right) \ , \tag{3.8}$$

see Table 1 for the definition of S and m_1. No expression for the power is available in the literature for the HB-FDR procedure.

4 Comparison of the AFD Procedure with the BH-FDR Procedure: A Simulation Study.

In our simulation, we consider one-sample multiple hypotheses testing

$$H_j : \ \mu_j = 0 \ \ v.s. \ \ \ A_j : \ \mu_j \neq 0, \ j = 1, \ldots, m. \tag{4.1}$$

We use the multivariate normal model $N_m(\boldsymbol{\mu}, \Sigma)$ to generate the data. We select $m = 1000$. Given m, we consider different values of the proportion m_0/m of the true cases and select $m_0/m = 0.2, 0.4, 0.6, 0.7, 0.8, 0.9$. For given m and m_0, let the first m_0 components follow the null hypotheses and have mean values $\mu_1, \ldots, \mu_{m_0} = 0$ and for the alternative cases, we generate the mean values $\mu_{m_0+1}, \ldots, \mu_m$ *iid* from Unif$(0.5, 1)$. For the covariance matrix Σ, we choose

$$\Sigma = (\rho_{ij}), \ \rho_{ij} = (0.8)^{|i-j|^{1/7}}. \tag{4.2}$$

For each combination of parameters, we simulate 10,000 datasets of sample size $N = 15$. For the AFD procedure, we have chosen $\delta = 0.005$ and for the BH-FDR procedure we have chosen $\alpha = 0.05$. For a fair comparison, the FDR of AFD is also controlled at level 0.05, assuming that T_i are independent and $m_0/m \geq 0.80$. From (4.2), it is clear that T_i's are neither independently distributed nor positively related and thus the procedure (3.7) should been used; we have, however, still used (3.6).

In Table 2, the estimated powers, FDR, and AFD of the two procedures, the BH-FDR and AFD are given along with their standard deviations below them. The standard deviations appear to be within acceptable limits although the standard deviation of the BH-FDR is always higher than the AFD. The power of the AFD procedure remains constant and have the same level for all the cases. The power of the BH-FDR procedure decreases as (m_0/m) increases. The power of the BH-FDR procedure is always higher than the AFD procedure for $m_0/m \leq 0.70$ but it has come at the cost of higher AFD. The BH-FDR procedure controls FDR at level $\alpha = 0.05$, but the entries show that it is only conservatively controlled for all the cases. This may be the reason for loss of power for this procedure. The FDR of the AFD procedure, although not controlled, is always below 0.05.

	mo/m	0.2	0.4	0.6	0.7	0.8	0.9
Power	AFD	0.391	0.393	0.394	0.391	0.391	0.391
	(se)	(0.119)	(0.122)	(0.125)	(0.128)	(0.131)	(0.141)
	BH-FDR	0.631	0.566	0.474	0.404	0.320	0.212
	(se)	(0.138)	(0.155)	(0.168)	(0.175)	(0.175)	(0.166)
AFD	AFD	0.005	0.005	0.005	0.005	0.005	0.005
	(se)	(0.024)	(0.026)	(0.024)	(0.021)	(0.019)	(0.020)
	BH-FDR	0.025	0.020	0.015	0.012	0.008	0.006
	(se)	(0.073)	(0.065)	(0.060)	(0.055)	(0.044)	(0.042)
FDR	AFD	0.002	0.006	0.013	0.018	0.027	0.050
	(se)	(0.011)	(0.025)	(0.043)	(0.057)	(0.075)	(0.118)
	BH-FDR	0.009	0.018	0.027	0.030	0.032	0.033
	(se)	(0.026)	(0.051)	(0.078)	(0.090)	(0.099)	(0.115)

Table 2: Attained Power, AFD, and FDR: $\delta = 0.005$, $\alpha = 0.054$, $m = 1000$. Correlation Structure as in (4.2)

From Table 2, it is clear that for $(\frac{m_0}{m}) \geq 0.7$, AFD performs better than the BH-FDR procedure.

Next in Table 3, we consider the case when the components are independently distributed. That is $\Sigma = \sigma^2 I$. In this case the FDR of the BH-FDR procedure can always be controlled at a specified level α which has been taken to be 0.05. The AFD for the AFD procedure is controlled at the level $\delta = 0.005$.

	mo/m	0.2	0.4	0.6	0.7	0.8	0.9
Power	AFD	0.393	0.392	0.393	0.393	0.392	0.393
	(se)	(0.017)	(0.020)	(0.024)	(0.028)	(0.035)	(0.049)
	BH-FDR	0.645	0.586	0.500	0.437	0.350	0.217
	(se)	(0.022)	(0.027)	(0.036)	(0.044)	(0.054)	(0.073)
AFD	AFD	0.005	0.005	0.005	0.005	0.005	0.005
	(se)	(0.005)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
	BH-FDR	0.026	0.018	0.010	0.007	0.004	0.001
	(se)	(0.012)	(0.007)	(0.004)	(0.003)	(0.002)	(0.001)
FDR	AFD	0.003	0.008	0.019	0.029	0.048	0.058
	(se)	(0.003)	(0.006)	(0.011)	(0.015)	(0.023)	(0.044)
	BH-FDR	0.010	0.020	0.030	0.035	0.040	0.045
	(se)	(0.004)	(0.007)	(0.012)	(0.016)	(0.023)	(0.045)

Table 3: The attained FDR of the AFD and BH-FDR; $\delta = 0.005$, $FDR = 0.05$. Independent Components. Here, FDR of AFD is also controlled at 0.05 for $m_0/m \geq 0.8$.

Again, from the Table 3, it is clear that for $(m_0/m) \geq 0.7$, AFD performs better than the BH-FDR procedure.

Finally, we consider the case when $\Sigma = \sigma^2[(1-\rho)I_m + \rho \mathbf{1}\mathbf{1}']$, that is, the components have intraclass correlation structure. Here I_m is an $m \times m$ identity matrix and $\mathbf{1}$ is an m-vector of ones, $\mathbf{1}' = (\mathbf{1}, \ldots, \mathbf{1})$. We choose $\rho = 0.5$ so that from Benjamini and Yekutieli (2001) , the FDR of the BH-FDR procedure can be controlled at the desired level. The simulation results are presented in Table 4, which shows that the AFD procedure still performs better than the BH-FDR procedure. Throughout the situation $\alpha = 0.05$ and $\delta = 0.005$.

	mo/m	0.2	0.4	0.6	0.7	0.8	0.9
Power	AFD	0.391	0.393	0.388	0.390	0.391	0.393
	(se)	(0.229)	(0.231)	(0.229)	(0.231)	(0.232)	(0.232)
	BH-FDR	0.601	0.547	0.461	0.408	0.388	0.234
	(se)	(0.282)	(0.288)	(0.290)	(0.288)	(0.273)	(0.242)
AFD	AFD	0.005	0.005	0.005	0.005	0.005	0.005
	(se)	(0.017)	(0.014)	(0.017)	(0.016)	(0.016)	(0.015)
	BH-FDR	0.022	0.017	0.012	0.009	0.007	0.004
	(se)	(0.053)	(0.044)	(0.040)	(0.036)	(0.034)	(0.030)
FDR	AFD	0.010	0.020	0.035	0.048	0.061	0.063
	(se)	(0.056)	(0.082)	(0.114)	(0.140)	(0.152)	(0.189)
	BH-FDR	0.010	0.018	0.027	0.032	0.033	0.035
	(se)	(0.041)	(0.062)	(0.090)	(0.105)	(0.112)	(0.127)

Table 4: The attained Power, AFD , and FDR for the AFD and BH-FDR. Intraclass Correlation Structure, correlation= 0.5; $\alpha = 0.05$, $\delta = 0.005$.

5 Examples from Microarrays

In this section, we use the multiple testing procedures discussed in previous sections to analyze two datasets from microarrays. The datasets are described next.

- **Colon Data:** This dataset, obtained by Affymetrix technology, contains the expression levels of 6500 genes, which were measured on 22 normal and 40 tumor colon tissues. From the original 6500 genes, Alon et al. (1999) has selected 2000 genes with the highest minimal intensity across the samples. Thus the selected dataset contains $p = 2000$ gene expression levels on $N_1 = 22$ normal subjects and $N_2 = 40$ tumor subjects.

- **Leukemia Data:** This dataset contains gene expression levels of 72 patients either suffering from acute lymphoblastic leukemia (ALL, 47 cases) or acute myeloid leukemia (AML 25 cases) and was obtained from Affymetrix oligonucleotide microarrays. More informationcan be found in Golub, et al. (1999). Following the protocol in Dudoit et al. (2002), we preprocess them by thresholding, filtering, a logarithmic transformation

and standardization, so that the data finally comprise the expression value of $p = 3571$ genes, and the degrees of freedom available for estimating the covariance is only 70.

The description of the above datasets and prepocessing are due to Dettling and Bühlmann (2002), except that we do not process the datasets such that each tissue sample has zero mean and unit variance across genes, which is not explainable in our framework. We roughly check the normality assumption by QQ-plotting around 50 genes selected randomly. The results are nearly satisfactory.

Before we embark on determining the number of differentially expressed genes by any of the methods, we need to check if the mean vectors of the two samples are indeed different. Also, to apply the BH-FDR procedure, we need to know if the statistics used for each hypothesis are independently distributed or at least positively related. We do these tasks in subsections 5.1, 5.2 and 5.3 respectively. In subsection 5.4, we obtain the differentially expressed genes by the BH-FDR and AFD methods and compare them with the lower bound given by Meinshausen and Bühlmann (2005) for these two data sets.

5.1 Testing the Equality of the Two Mean Vectors.

Denote the mean vectors of the two samples based on N_1 and N_2 observation vectors $\boldsymbol{x}_{ij}, j = 1, \ldots, N_i, i = 1, 2$, by

$$\overline{\boldsymbol{x}}_1 = N_1^{-1} \sum_{j=1}^{N_1} \boldsymbol{x}_{1j} \ , \quad \overline{\boldsymbol{x}}_2 = N_2^{-1} \sum_{j=1}^{N_2} \boldsymbol{x}_{2j} \ ,$$

and the pooled sample covariance matrix by

$$\begin{aligned} S &= n^{-1} \sum_{i=1}^{2} \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i) (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)' \ , \ n = N_1 + N_2 - 2 \\ &= (s_{ij}) \ , \end{aligned}$$

Let $D_s = \mathrm{diag}(s_{11}, \ldots, s_{mm})$, be the $m \times m$ diagonal matrix and

$$R = D_s^{-\frac{1}{2}} S D_s^{-\frac{1}{2}}$$

be the correlation matrix. For testing the equality of the mean vectors of the two samples, we use the statistic, due to Srivastava and Du (2008),

$$T = \frac{q(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' D_s^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (n/n-2)m}{[2(\mathrm{tr}R^2 - m^2/n) c_{m,n}]^{\frac{1}{2}}},$$

where

$$q = N_1 N_2 / (N_1 + N_2), \text{ and } c_{m,n} = 1 + \mathrm{tr}R^2 / m^{\frac{3}{2}} \ .$$

Under the hypothesis that the two mean vectors are equal, it is asymptotically distributed as $N(0, 1)$. The values of the statistics T are 4.6882 and 17.0758 respectively. Hence, the two mean vectors are not equal in both the examples. We also used the statistic proposed by Srivastava (2007) and obtained the same result.

5.2 Test for Independence.

In order to test for the independence of the t-statistics, we need to test the hypothesis that the population covariance matrix is a diagonal matrix. For this, Srivastava (2005, 2006) proposed two test statistics. One of them is given by

$$Q = \frac{(n+2)\sum_{i<j} z_{ij}^2 - \frac{1}{2}m(m-1)}{\sqrt{m(m-1)}} \rightarrow N(0,1),$$

where

$$z_{ij} = \frac{1}{2}\log\frac{1+r_{ij}}{1-r_{ij}}\ ,\ i \neq j,\ i,j = 1,\ldots,m\ .$$

Under the hypothesis of independence, the test statistics Q is $N(0,1)$. The p-values are zero for both data sets; the second test, not presented here, also gives zero p-values. Hence, the hypothesis of diagonality for both datasets is rejected.

5.3 Test for Positively Relatedness of the Statistics for Testing the Hypotheses, $H_i, i = 1,\ldots,m$

Benjamini and Yekutieli (2001) have shown that the student's t-statistics will be positively related if the covariance matrix is of intraclass correlation structure with positive correlation. Applying an orthogonal transformation of Helmert's type, Srivastava (2006) showed that it is equivalent to testing the spherecity of an $(m-1)\times(m-1)$ matrix. The p-value of this test is also zero which implies that for both datasets the test statistics may not be positively related.

5.4 Differentially Expressed Genes for the Two Data Sets

From the results of sections 5.2 and 5.3, it is clear that the control of the FDR for the BH-FDR procedure cannot be guaranteed. Using the two-sample student's t-statistics T_j, we obtain

$$p_j = \mathrm{P}\{|T_j| \geq |t_j|\},\ \ j = 1,\ldots,m,$$

where the probability depends on both the set Λ_0 of true null hypotheses and the correlation structure of T_j's. We then order the obtained p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.

Using nominal significane level $\alpha = 0.05$, we have the following results shown in Table 3.

The AFD numbers are closer to the numbers given by Meinshausen and Bühlmann (2005). In any case, the AFD procedure appears to be a viable procedure in determining the number of differentially expressed genes. In fact, since the T_i's are neither independently distributed nor positively related, the procedure in (3.7) should have been used.

Colon:	m	2000	Leukemia:	m	3571
	BH-FDR	354		BH-FDR	1105
	AFD	292		AFD	865
	MB	286		MB	957

Table 5: First two rows show the Numbers of differentially expressed genes selected by each AFD and BH-FDR procedure for colon and leukemia datasets, respectively. Nominal significance level $\alpha = 0.05$; $\delta = 0.005$. The last row is the lower bound given in Meinshausen and Bühlmann (2005) at level 0.05.

6 Conclusion

We observed that both the BH-FDR and the AFD procedures control the expected values of some unobservable random variables. When $m_0 = m$, the FDR procedure becomes FWER procedure. Similarly, when $m_0 = m$, the AFD procedure becomes PCER procedure. The BH-FDR can be controlled at the specified level if the statistics T_i's used in testing the hypotheses H_i's are independently distributed or positively related. The Attained Significance Level (ASL) however, depends om m_1 and m. For large m_1, the ASL of the BH-FDR procedure has been found to be much smaller than the specified level α, and thus resulting in loss of power. On the other hand, the AFD can be controlled irrespective of how these T_i's are related and the values of m_1 and m. The power of the BH-FDR method, depends not only on the alternative hypotheses or the false hypotheses but also on the number of the false hypotheses m_1 as well as on the total number of hypotheses m. This is not the case with the AFD procedure. In fact, if in the alternative all the false hypotheses have the same distribution, the power of the AFD procedure remains constant irrespective of the values of m_0, m_1 and m. In power comparison, we notice that the BH-FDR has higher power than the AFD procedure when the ratio $m_0/m \leq 0.70$ at the cost of higher AFD, but have always smaller power when $m_0/m > 0.70$ and AFD above the specified limit. It has been observed in practice that in most practical applications, m_0/m may even be larger than 0.9. The FDR of the BH-FDR procedures is controlled at level $\alpha = 0.05$, and the AFD of the AFD procedure is controlled at level $\delta = 0.005$, which can be chosen to control k-FWER, γFWER or even FDR assuming independence of T_i's. In terms of powers, the AFD procedure has higher power, and lower AFD than the BH-FDR procedure for large values of m_0/m. Thus, we may conclude that the AFD procedure, which has a constant critical region, is a well suited procedure in practical applications such as in microarrays. The two examples also support this conclusion.

Acknowledgement

This research was partially supported by Natural Sciences and Engineering Research Council of Canada.

Appendix: Proofs

In this section, we will give the proofs of Lemma 2.1 and 2.2 and of Corollary 2.1 and 2.2. For simplicity, we shall denote $I(P_j \leq \delta | H_j \in \Lambda_0)$ by $I_{\Lambda_0}(P_j \leq \delta)$ and $\mathrm{P}(P_j \leq \delta | H_j \in \Lambda_0)$ by $\mathrm{P}_{\Lambda_0}(P_j \leq \delta)$.

Proof of Lemma 2.1. We have

$$\begin{aligned}
\mathrm{Var}(V/m_0) &= \mathrm{E}[V/m_0]^2 - (\mathrm{E}[V/m_0])^2 = \frac{1}{m_0^2}\mathrm{E}_{\Lambda_0}[\sum_{j\in\Lambda_0} I(P_j \leq \delta)]^2 - \delta^2 \\
&= \frac{1}{m_0^2}\mathrm{E}_{\Lambda_0}[\sum_{j\in\Lambda_0} I(P_j \leq \delta) + \sum_{j,l\in\Lambda_0, j\neq l} I(P_j \leq \delta) I(P_l \leq \delta)] - \delta^2 \\
&= \frac{\delta}{m_0} + \frac{1}{m_0^2}\sum_{j,l\in\Lambda_0, j\neq l} \mathrm{E}_{\Lambda_0}[I(P_j \leq \delta) I(P_l \leq \delta)] - \delta^2. \qquad \text{(A.1)}
\end{aligned}$$

Proof of Lemma 2.2. For large N, G may be considered as normal distribution Φ. Then for large N,

$$\begin{aligned}
\psi(\delta, \rho_{jl}) &= \mathrm{E}_{\Lambda_0}[I(P_j \leq \delta, P_l \leq \delta)] = \mathrm{P}[|T_j| \geq t_0, |T_l| \geq t_0] \\
&\simeq \mathrm{P}[|Z_j| \geq t_0, |Z_l| \geq t_0] = \mathrm{P}[|Z_j| \geq t_0, Z_l \geq t_0, or\ Z_l \leq -t_0], \qquad \text{(A.2)}
\end{aligned}$$

where $Z_j, Z_l \sim N(0,1)$ and $\mathrm{Cov}(Z_j, Z_l) = \rho_{jl}$. Hence,

$$\begin{aligned}
\psi(\delta, \rho_{jl}) &= \mathrm{P}\left(|Z_j| \geq t_0, \frac{Z_l - \rho_{jl}Z_j}{\sqrt{1-\rho_{jl}^2}} \geq \frac{t_0 - \rho_{jl}Z_j}{\sqrt{1-\rho_{jl}^2}}, or\ , \frac{Z_l - \rho_{jl}Z_j}{\sqrt{1-\rho_{jl}^2}} \leq \frac{-t_0 - \rho_{jl}Z_j}{\sqrt{1-\rho_{jl}^2}}\right) \\
&= \frac{1}{\sqrt{2\pi}}\int_{|x|\geq t_0}\left[\Phi\left(\frac{-t_0 + \rho_{jl}x}{\sqrt{1-\rho_{jl}^2}}\right) + \Phi\left(\frac{-t_0 - \rho_{jl}x}{\sqrt{1-\rho_{jl}^2}}\right)\right]\exp(-x^2/2)dx. \qquad \text{(A.3)}
\end{aligned}$$

Proof of Corollary 2.1. Since P_j's are independently distributed, we get from (A.1) that

$$\mathrm{E}_{\Lambda_0}[I(P_j < \delta) I(P_l < \delta)] = \mathrm{P}_{\Lambda_0}(P_j < \delta)\mathrm{P}_{\Lambda_0}(P_l < \delta) = \delta^2. \qquad \text{(A.4)}$$

Hence,

$$\mathrm{Var}(V/m_0) = \frac{\delta}{m_0} + \frac{m_0(m_0-1)}{m_0^2}\delta^2 - \delta^2 = \delta(1-\delta)/m_0. \qquad \text{(A.5)}$$

References

[1] Alon, U., Barkai, N., Motterman, D., Gish, K., Mack, S., and Levine, J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by olingonucleotide. PNAS, 6745–6750.

[2] 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, Series B*, **57**, 289–300.

[3] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1181.

[4] Black, M. A. (2004). A note on the adaptive control of false discovery rate. *Journal of the Royal Statistical Society, Series B*, **66**, 297–304.

[5] Dettling, M. and Bühlmann, P. (2002). Boosting for tumor classification with gene expression data. *Bioinfomatics*, 1–9.

[6] Du, M. and Srivastava, M. S. (2006). Comparison of multiple testing procedures and the analysis of two examples from microarrays. *JSM Proceedings*, 271-222.

[7] Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

[8] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of Royal Statistical Society, Series B*, **64**, 499–517.

[9] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaassenbeek, M., Mesirov, J., Coller, H., Loh, M., and Downing, J. (1999). Molecular classification of cancer class discovery and class prediction by gene expression monitoring. *Science*, 531–537.

[10] Hommel, G. and Hoffman, T. (1988). Controlled uncertainty. In Multiple Hypothesis Testing. P. Bauer, G. Hommel, and Sonnenmann, eds., 154–161. Springer, Heidelberg.

[11] Lehmann, E. L. and Romano, J. P. (2005). Generalization of familywise error rate. *Annals of Statistics*, **33**, 1138–1154.

[12] Meinshausen and Bühlmann, P. (2005). Lower bounds for the number of false null hypothesis for multiple testing of associations under general dependence structures. *Biometrika*, **92**, 893–907

[13] Romano, J. P., and Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics*, **34**, 1850–1873.

[14] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754

[15] Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *Journal of Japan Statistical Society*, **35**, 251–272

[16] Srivastava, M. S. (2006). Some tests criteria for the covariance matrix with fewer observations than the dimension. *Acta Et Commentationes Universitatis Tartuensis De Mathematica*, **10**, 77–93.

[17] Srivastava, M. S. (2007). Multivariate theory for analyzing high-dimensional data. *Journal of Japan Statistical Society*, **37**, 53–86

[18] Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal Multivariate Analysis*, **99**(3), 386–402.