Biometrika (2006), **93,** 3, *pp.* 509–524

Printed in Great Britain

False discovery control with *p*-value weighting

By CHRISTOPHER R. GENOVESE, KATHRYN ROEDER
and LARRY WASSERMAN

Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

genovese@stat.cmu.edu roeder@stat.cmu.edu larry@stat.cmu.edu

Summary

We present a method for multiple hypothesis testing that maintains control of the false discovery rate while incorporating prior information about the hypotheses. The prior information takes the form of *p*-value weights. If the assignment of weights is positively associated with the null hypotheses being false, the procedure improves power, except in cases where power is already near one. Even if the assignment of weights is poor, power is only reduced slightly, as long as the weights are not too large. We also provide a similar method for controlling false discovery exceedance.

Some key words: False discovery rate; Multiple testing; Power.

1. Introduction

Data from DNA microarray experiments, genetic epidemiology, functional Magnetic Resonance Imaging (fMRI) experiments and astronomical imaging have spurred renewed interest in the multiple testing problem of controlling overall error rates when performing simultaneous hypothesis tests. These datasets share two features that distinguish them from multiple-testing examples in traditional linear models. First, there are often many thousands, even millions, of null hypotheses to test. This exacerbates the trade-off between power and Type I error control, making it more difficult to detect small effects, which are often of the greatest interest. Secondly, the tests are related by a scientifically meaningful structure: in fMRI, each test corresponds to a specific brain location; in microarray studies, each test corresponds to a specific gene. Both scientific and spatial prior information can thus be exploited to improve the performance of testing procedures. Put simply, all null hypotheses are not created equal.

The introduction of the false discovery rate and a procedure to control it by Benjamini & Hochberg (1995) gave an effective way to address the first issue above. Control of false discovery rate lets investigators increase power while maintaining a principled bound on error. The procedure of Benjamini & Hochberg (1995), to which we refer as the BH procedure, is fast and easy even with large datasets and performs well in sparse cases where there are relatively few true alternatives (Genovese & Wasserman, 2002a). Let $P_{(1)} < \ldots < P_{(m)}$ be the ordered *p*-values from m hypothesis tests, with $P_{(0)} \equiv 0$. Then the

BH procedure rejects any null hypothesis for which $P \leqslant T$ with

$$T = \max\left\{P_{(i)} : P_{(i)} \leqslant \frac{\alpha i}{m}\right\};$$

this controls the false discovery rate at level $\alpha m_0/m$, where m_0 is the number of true null hypotheses. Adaptive variants of the procedure can increase power further at little additional computational expense; see Benjamini et al. (2006) and Storey (2002).

However, neither this procedure nor its variants deal with the issue of structure and prior information, because they treat all null hypotheses interchangeably. For example, previous studies can suggest that some null hypotheses are more, or less, likely to be false. Similarly, in spatial problems, false nulls are more likely to be clustered than true nulls. In this paper, we consider the approach of expressing prior information through weights on each null hypothesis.

The idea of weighting hypotheses is not new. We distinguish two approaches, namely *p*-value weighting, as above, and loss weighting, in which each weight is placed on the loss or error criterion for the corresponding incorrect decision. Holm (1979) introduced the idea of *p*-value weights. A larger weight can be used to suggest that it is more likely that the null hypothesis is false a priori. Holm (1979) showed that his sequential step-down test maintains control of the familywise error rate when the *p*-values are divided by weights, as long as the step-down constants are adjusted appropriately. Benjamini & Hochberg (1997) investigated the use of weighting in a variety of settings. They used weights in the definition of the error rate, i.e. loss weighting, to indicate the importance of each hypothesis. Here, we use *p*-value weighting as a frequentist method for including prior information about the hypotheses, leaving the error measure unchanged.

Such prior information is often available in practice. In fMRI studies, for example, information about the functional response to a stimulus can be gleaned from previous studies, pilot data and direct neural recording in animals. Detailed anatomical information is also available from structural images of each subject. Similarly, in genetic epidemiology tens of thousands of genomic regions may be tested in a genetic association study to locate alleles that increase the risk for complex diseases. Frequently the association study is conducted after genetic linkage studies have been published. In contrast to an association study, which is designed to pinpoint genetic variants associated with disease, linkage analysis points to very broad regions of the genome that appear to contain genetic variants of interest. These regions often contain tens or even hundreds of genes. Initially it might seem that such information would not be refined enough to offer reliable weights; however, coupled with the partial knowledge of genetic function available from the human genome project, linkage studies are likely to provide useful guidance for choosing weights in an association study.

In general, *p*-value weighting raises several important questions. How can we choose weights so as to maintain control of a suitable error criterion, such as the false discovery rate? How much power can we gain if we guess well in the weight assignment? How much power can we lose if we guess poorly? In this paper, we will present a *p*-value weighting procedure that controls the false discovery rate. We will show that, under moderately informative guessing, weighting improves power nontrivially and that, under even misinformative guessing, the worst-case loss in power is small. We also explore the role of weights when controlling false discovery exceedance; see Genovese & Wasserman

(2004), Korn et al. (2004), an unpublished Carnegie Mellon University technical report by C. R. Genovese and L. Wasserman and an unpublished technical report by M. van der Laan, S. Dudoit and K. S. Pollard from the University of California at Berkeley. The reader interested primarily in our procedures need read only § 2 for notation, § 3 for control of the false discovery rate and § 5 for control of the false discovery exceedance.

2. The *p*-value weighting

Consider the simplest case where, based on previous studies and results, an investigator can partition the m null hypotheses into two groups, where the null is a priori more plausible in half of the hypotheses and the alternative in the others. In this setting, it seems reasonable to consider using different thresholds for hypotheses in each of the two groups. If the (random) thresholds are T_0 and T_1, say, and if $j(i)$ is the group for the ith null hypothesis, we can write $P_i \leqslant T_{j(i)}$ as $P_i \leqslant W_i T$, or equivalently $P_i/W_i \leqslant T$, where $T = (T_0 + T_1)/2$ and $W_i = 2T_{j(i)}/(T_0 + T_1)$. Thus, using different thresholds for the groups corresponds to using a single threshold but weighting the p-values. Note that, when $T_0 \neq T_1$, the weights will be bigger than 1 in one group and less than 1 in the other. A binary weighting scheme is just one choice among the much broader class of possibilities for weights.

Whatever information one uses to construct p-value weights, the weight assignment remains a guess. This guess is to be made a priori, that is before seeing the p-values. For the purposes of analysis, we model the weights as random variables that are related to the underlying truth or falsehood of each null hypothesis.

Let $P^m = (P_1, \ldots, P_m)$ denote the observed p-values, with $P_{(1)} < \ldots < P_{(m)}$ denoting the ordered p-values and $P_{(0)} \equiv 0$. Define hypothesis indicator variables $H^m = (H_1, \ldots, H_m)$, where $H_i = 0$ (or $= 1$) if the ith null hypothesis is true (or false). Let the p-value weights be random variables $W^m = (W_1, \ldots, W_m)$ that are conditionally independent of P^m given H^m.

We will assume that the p-values are drawn from the following mixture model: $H_i \sim \mathrm{Ber}(a)$; $P_i|H_i = 0 \sim \mathrm{Un}(0, 1)$; and $P_i|H_i = 1 \sim F$, where $0 < a < 1$ and F is a probability distribution on $(0, 1)$ which we assume is stochastically smaller than the Uniform. Marginally, the p-values are drawn from the probability distribution $G = (1 - a)U + aF$, where U is the probability distribution of a $\mathrm{Un}(0, 1)$.

We consider two models for generating W^m. The first is a special case that makes analysis more concrete; the second is essentially general. For both we assume that the weights satisfy $\bar{W}_m \equiv (1/m) \sum W_i = 1$.

Case 1: *Binary weighting*. In this case, the weights W^m can take on two possible values $w_0 \leqslant 1 \leqslant w_1$. To derive W^m, let $U^m = (U_1, \ldots, U_m)$ be $\mathrm{Ber}(\gamma)$ random variables representing prior guesses for each of the hypotheses, with $U = 1$ corresponding to an alternative and $U = 0$ to a null. Let $\bar{U}_m = m^{-1} \sum_{i=1}^m U_i$ be the fraction bet on the alternative. Assume that U^m and P^m are conditionally independent given H^m and that, given H^m, the (P_i, U_i) pairs are independent.

The U_i's are related to the H_i's through $\eta = \mathrm{pr}(U = 1|H = 1)/\mathrm{pr}(U = 1|H = 0)$, a measure of the informativeness of guessing. When $\eta = 1$, U^m and H^m are independent; for $\eta > 1$, there is greater likelihood of guessing correctly, and, for $0 \leqslant \eta < 1$, incorrectly. To be

specific, we assume that each U_i is drawn marginally from a Ber(γ) with conditional distribution

$$\mathrm{pr}(U_i = 1|H_i = 1) = \frac{\gamma\eta}{a\eta + 1 - a}, \quad \mathrm{pr}(U_i = 1|H_i = 0) = \frac{\gamma}{a\eta + 1 - a}.$$

To indicate the strength of the weights we use the parameter r in our definition of the weights based on the U_i's as follows:

$$W_i = \frac{1 + (r-1)U_i}{1 + (r-1)\bar{U}_m}. \tag{1}$$

A weight in favour of the null takes on value $w_0 = 1/\{1 + (r-1)\bar{U}_m\}$, and likewise a weight in favour of the alternative is $w_1 = r/\{1 + (r-1)\bar{U}_m\}$. Note that $r = w_1/w_0$ and, when the parameter r is 1, we return to the unweighted case. As desired, the mean weight satisfies $\bar{W}_m = 1$.

In summary, the proposed weighting scheme has three parameters which are interpreted as follows: γ determines the prevalence of alternative guesses; η determines the informativeness of guessing; and r determines the strength of weighting.

Case 2: *General weighting*. More generally, let $U_1, \ldots, U_m$ be random variables with sample mean $\bar{U}_m$ and define weights $W_i = U_i/\bar{U}_m$. These weights satisfy $\mathrm{pr}(\bar{W}_m = 1) = 1$ by construction. The W_i's are not independent, but they are, marginally, identically distributed with $W_i|H_i = 0 \sim M_0$ and $W_i|H_i = 1 \sim M_1$, for probability distributions M_0 and M_1 on $[0, \infty)$. The marginal distribution of W is then $M = (1-a)M_0 + aM_1$. For $j = 0, 1$, let $\mu_j = E(W|H = j)$, the means of M_0 and M_1 respectively, and let $\mu = (1-a)\mu_0 + a\mu_1$ be the marginal mean, which equals 1 by design.

3. CONTROL OF FALSE DISCOVERY RATE WITH PRIOR WEIGHTING

If we reject all hypotheses for which $P_i \leqslant T$, for some, possibly random, threshold T, then the false discovery proportion is defined to be

$$\mathrm{FDP}(T) = \frac{\text{false rejections}}{\text{rejections}} = \frac{\sum_{i=1}^n 1\{P_i \leqslant T\}(1 - H_i)}{\sum_{i=1}^n 1\{P_i \leqslant T\}},$$

where the ratio is defined to be zero when the denominator is zero. For threshold T, the false discovery rate is defined to be $E\{\mathrm{FDP}(T)\}$.

We now define a procedure, which we call WBH, for incorporating prior p-value weights while maintaining control of false discovery rate. Following Holm (1979), we weight the p-values using prior weights W_i. Define $Q_i = P_i/W_i$, where $W_i \geqslant 0$. In practice the weights adjust the threshold for rejection: rejecting when $Q_i \leqslant T$ is equivalent to rejecting when $P_i \leqslant W_i T$.

Several authors have noted that the threshold for the BH procedure can be written as

$$T_{\mathrm{BH}} = \sup\{t : \hat{B}(t) \leqslant \alpha\},$$

where $\hat{B}(t) = t/\hat{G}_m(t)$ (Benjamini & Hochberg, 1995; Storey, 2002; Genovese & Wasserman, 2002a). Thus the BH procedure is based upon $\hat{G}_m(t)$ which estimates $G(t) = \mathrm{pr}(P \leqslant t)$. Analogously the WBH procedure can be based upon $D(t) = \mathrm{pr}(P/W \leqslant t)$, which equals

$$\begin{aligned} D(t) &= \int \mathrm{pr}\left(\frac{P}{W} \leqslant t | W = w\right) dM(w) = \int \sum_{h=0}^{1} \mathrm{pr}\left(\frac{P}{W} \leqslant t | W = w, H = h\right) f(h|w) dM(w) \\ &= \int \sum_{h=0}^{1} \mathrm{pr}(P \leqslant wt | H = h) f(h|w) dM(w) = \int \sum_{h=0}^{1} \{(1-h)tw + hF(tw)\} f(h|w) dM(w) \\ &= \sum_{h=0}^{1} \int \{(1-h)tw + hF(tw)\} dM(w|h) f(h) \\ &= (1-a) \int twdM(w|h=0) f(h) + a \int F(tw) dM(w|h=1) \\ &= (1-a)\mu_0 t + a \int F(wt) dM_1(w). \end{aligned} \tag{2}$$

By analogy with the BH procedure we incorporate the weights by defining

$$T_{\mathrm{WBH}} = \sup\{t : \hat{R}(t) \leqslant \alpha\}, \tag{3}$$

where

$$\hat{R}(t) = \frac{t \sum_{i=1}^{m} W_i}{\sum_{i=1}^{m} 1\{P_i \leqslant W_i t\}} = \frac{t}{\hat{D}_m(t)},$$

and $\hat{D}_m(t)$ is the empirical probability distribution of P_i/W_i.

The procedure is as follows.

Step 1. Assign weights $W_i > 0$ to each null hypothesis such that $\bar{W}_m = 1$.

Step 2. For each $i = 1, \ldots, m$, compute $Q_i = P_i/W_i$.

Step 3. Apply the BH procedure at level α to the Q_i's.

In § 4, we show that the WBH controls the false discovery rate at the nominal level. In §§ 6 and 7, we investigate the power of the procedure.

4. The WBH controls false discovery rate

We begin with a finite-sample result and then describe the asymptotic behaviour of the WBH threshold.

Theorem 1. (i) *Conditionally on* H^m, WBH *controls the false discovery rate at the level* $\alpha\mu_0 m^{-1} \sum_i (1 - H_i)$. (ii) *Unconditionally* WBH *controls the false discovery rate at the level* $\alpha(1-a)\mu_0 \leqslant \alpha$.

Proof. Our approach is based on the method of Benjamini & Yekutieli (2001). First, let $Q_i = P_i/W_i$ with sorted values $Q_{(i)}$ as usual. Note that the WBH threshold (3) for the Q_i's can be equivalently written as $T = \sup\{Q_{(i)} : Q_{(i)} \leqslant \alpha i/m\}$.

Let $q_k = \alpha k/m$. Define the events

$$R_{k,i} = \left\{ \sum_{j \neq i} 1\{Q_j \leqslant q_k\} = k-1 \right\}.$$

For each i, the events $R_{k,i}$ for $k = 1, \ldots, m$ form a partition: they are disjoint and $\cup_{k=1}^{m} R_{k,i}$ must occur. To see the last point, note that for each realisation $k \mapsto 1 + \sum_{j \neq i} 1\{Q_j \leqslant q_k\}$ is a nondecreasing function from $\{1, \ldots, m\}$ into $\{1, \ldots, m\}$ and thus has a fixed point. Then

$$\begin{aligned}
E\{\mathrm{FDP}(T)|H^m\} &= \sum_{i:\, H_i=0} \sum_{k=1}^{m} \frac{1}{k} \operatorname{pr}(\{P_i \leqslant W_i q_k\} \cap R_{k,i} | H_i = 0, H^m) \\
&= \sum_{i:\, H_i=0} \sum_{k=1}^{m} \frac{1}{k} E\{\operatorname{pr}(\{P_i \leqslant W_i q_k\} \cap R_{k,i} | H_i = 0, H^m, W^m) | H_i = 0, H^m\} \\
&= \sum_{i:\, H_i=0} E\left\{ \sum_{k=1}^{m} \frac{1}{k} \operatorname{pr}(P_i \leqslant W_i q_k | H_i = 0, H^m, W^m) \right. \\
&\qquad \left. \times \operatorname{pr}(R_{k,i} | H^m, W^m) | H^m, H_i = 0 \right\} \qquad (4) \\
&= \sum_{i:\, H_i=0} E\left\{ \sum_{k=1}^{m} \frac{\alpha W_i}{\sum_j W_j} \operatorname{pr}(R_{k,i} | H^m, W^m) | H^m, H_i = 0 \right\} \qquad (5) \\
&= \sum_{i:\, H_i=0} E\left\{ \frac{\alpha W_i}{m} \sum_{k=1}^{m} \operatorname{pr}(R_{k,i} | H^m, W^m) | H^m, H_i = 0 \right\} \\
&= \sum_{i:\, H_i=0} E\left(\frac{\alpha W_i}{m} | H_i = 0, H^m \right).
\end{aligned}$$

Equation (4) follows from the conditional independence of P_i and the remaining p-values. In the case of discrete test statistics, equality in (5) need not hold, but it can be replaced by a '$\leqslant$' as mentioned by Benjamini & Yekutieli (2001).

It follows that

$$E\{\mathrm{FDP}(T)|H^m\} = \alpha E(W_1 | H_1 = 0) \frac{1}{m} \sum_i (1 - H_i) = \alpha \mu_0 \frac{1}{m} \sum_i (1 - H_i).$$

This proves part (i) of the Theorem. Taking expectations under the mixture model produces $\alpha\mu_0(1-a)$ on the right-hand side. For $\mu = 1$, $\mu_0 \leqslant 1/(1-a)$ and this bound is $\leqslant \alpha$ as stated in part (ii). □

Next we explore the asymptotic behaviour of the procedure. The resulting theoretical developments support the finite sample theorem and also pave the way for derivations of more general results that do not assume the p-values to be independent and identically distributed.

Define $C(t) = D(t)/t$, $\hat{C}_m(t) = \hat{D}_m(t)/t$, and corresponding threshold

$$t_* = \sup\left\{ t : C(t) \geqslant \frac{1}{\alpha} \right\}, \quad T_m = \sup\left\{ t : \hat{C}_m(t) \geqslant \frac{1}{\alpha} \right\}.$$

We then have the following result.

LEMMA 1. *If F is strictly concave on $[0,1]$, then* (i) *G is strictly concave on $[0,1]$,* (ii) *D is strictly concave on $[0,1]$, and* (iii) *C is monotone decreasing on $(0,1)$.*

Proof. Since $G=(1-a)U+aF$, the first claim follows immediately. Similarly, by equation (2), $D(t)=(1-a)\mu_0 t+a\int F(wt)dM(w|h=1)$. Hence, for $0\leqslant\lambda\leqslant 1$,

$$D\{(1-\lambda)t_0+\lambda t_1\}=(1-a)\mu_0\{(1-\lambda)t_0+\lambda t_1\}+a\int F[w\{(1-\lambda)t_0+\lambda t_1\}]dM(w|h=1)$$
$$>(1-\lambda)D(t_0)+\lambda D(t_1),$$

where the final inequality is strict because $M(.|h=1)$ has all its mass on $[0,\infty)$. This proves part (ii).

Finally, let $1>t_1>t_0>0$ and note that $F(0)=0$ implies that $D(0)=0$. Note that

$$C(t_1)=\frac{D(t_1)}{t_1}=\frac{(1-t_0/t_1)D(0)+(t_0/t_1)D(t_1)}{t_0}\leqslant\frac{D(t_0)}{t_0}=C(t_0),$$

which proves part (iii). □

THEOREM 2. *Assume that F is strictly concave. Then $T_m\to t_*$ almost surely,*

$$E|\mathrm{FDP}(T_m)-\mathrm{FDP}(t_*)|\to 0,$$

and thus $E\{\mathrm{FDP}(T_m)\}\leqslant\alpha+o(1)$.

A proof is provided in the Appendix.

5. WEIGHTED EXCEEDANCE CONTROL

In this section, we present an approach to weighted testing that controls false discovery exceedance. The method generalises the approaches of Genovese & Wasserman (2004), Perone Pacifico et al. (2004) and M. van der Laan in an unpublished report. Those methods begin with a familywise test and then augment the familywise rejection region by adding in extra rejections.

Let $S=\{1,\ldots,m\}$ and let $S_0\equiv\{j:H_j=0\}$ be the set of the true nulls. We call any, possibly random, $\mathscr{R}\subset S$ a rejection region and say that $\mathscr{R}$ controls familywise error rate at level α if $\mathrm{pr}\{\#(\mathscr{R}\cap S_0)>0\}\leqslant\alpha$, where $\#(B)$ denotes the number of points in a set B. The false discovery proportion of a rejection set $\mathscr{R}$ is then

$$\mathrm{FDP}=\frac{\sum_{j=1}^m(1-H_j)1\{j\in\mathscr{R}\}}{\sum_{j=1}^m 1\{j\in\mathscr{R}\}},$$

where the ratio is defined to be zero if the denominator is zero.

Instead of controlling the mean of the false discovery rate, we will instead control the false discovery proportion exceedance. To be specific, our goal in this section is to use the weighted p-values to find a rejection set $\mathscr{R}$ such that

$$\mathrm{FDX}\equiv\mathrm{pr}(\mathrm{FDP}>c)\leqslant\alpha, \tag{6}$$

for given c and α. We call such an $\mathscr{R}$ a (c, α) rejection region and we say that $\mathscr{R}$ provides (c, α) exceedance control. The inequality (6) will be valid for all finite m and will not make assumptions about the form of the dependence between the p-values.

We begin by introducing weighted familywise tests. Then we use these familywise tests to construct exceedance control methods. We recall two popular methods for familywise control. Recall that $P_{(1)} \leqslant \ldots \leqslant P_{(m)}$ denote the sorted p-values. The Bonferroni method uses $\mathscr{R}_0 = \{j : P_j \leqslant \alpha/m\}$. Holm's (1979) method takes $\mathscr{R}_0 = \{j : P_j \leqslant T\}$, where $T = 0$ if $P_{(1)} \geqslant \alpha/m$ and $T = P_{(k)}$ otherwise, where $k = \max\{j : P_{(j)} < \alpha/(m-j)\}$.

The weighted Bonferroni rejection set is $\mathscr{R}_0 = \{j : Q_j \leqslant \alpha/m\}$.

LEMMA 2. *The weighted Bonferroni procedure controls familywise error at level* $(1-a)\mu_0\alpha \leqslant \alpha$.

Proof. We have that

$$\begin{aligned}
\operatorname{pr}\{\#(\mathscr{R} \cap S_0) > 0\} &= \operatorname{pr}\left(P_j \leqslant \frac{\alpha W_j}{m} \text{ for some } j \in S_0\right) \leqslant \sum_{j=1}^{m} \operatorname{pr}\left(P_j \leqslant \frac{\alpha W_j}{m} \text{ and } H_j = 0\right) \\
&= \sum_{j=1}^{m} \operatorname{pr}\left(P_j \leqslant \frac{\alpha W_j}{m} | H_j = 0\right) \operatorname{pr}(H_j = 0) \\
&= (1-a) \sum_{j=1}^{m} \int \operatorname{pr}\left(P_j \leqslant \frac{\alpha w_j}{m} | H_j = 0, W^m = w^m\right) dM(w^m | h_j = 0) \\
&= (1-a)\frac{\alpha}{m} \sum_{j=1}^{m} \int w dM(w^m | h_j = 0) = (1-a)\alpha\mu_0 \leqslant \alpha. \qquad \square
\end{aligned}$$

Holm's (1979) weighted procedure for controlling familywise error is as follows. Let $Q_i = P_i/W_i$ and let $Q_{(1)} \leqslant \ldots \leqslant Q_{(m)}$ denote the sorted values. Let $H_{(1)}, \ldots, H_{(m)}$ and let $W_{(1)}, \ldots, W_{(m)}$ denote the H_i's and W_i's sorted correspondingly. If $Q_{(1)} \geqslant \alpha/m$, set $\mathscr{R}_0 = \varnothing$; otherwise, find the largest i for which

$$Q_{(j)} \leqslant \frac{\alpha}{\sum_{k=j}^{m} W_{(j)}}$$

holds for all $j = 1, 2, \ldots, i$. Let $\mathscr{R}_0$ be the indices corresponding to the j smallest Q_j's. Holm proved that

$$\operatorname{pr}\{\#(\mathscr{R}_0 \cap S_0) > 0\} \leqslant \alpha$$

when the weights and H_i's are regarded as fixed. We now prove that the same is true for random weights and random H_i's by adapting his proof.

LEMMA 3. *The weighted Holm procedure controls familywise error at level* α.

Proof. Let $H^m = (H_1, \ldots, H_m)$, $W^m = (W_1, \ldots, W_m)$, $N_0 = \sum_{i=1}^{m} W_i(1-H_i)$ and let $\pi(h)$ denote the marginal probability mass function for the vector H^m. Also, we write $S_0 = S_0(H^m)$ to make explicit the dependence of S_0 on H^m. Define the event

$$A = \left\{Q_i > \frac{\alpha}{N_0} \text{ for all } i \in S_0\right\}.$$

Then

$$\begin{aligned}
\mathrm{pr}(A) &= 1 - \mathrm{pr}\left(Q_i \leqslant \frac{\alpha}{N_0} \text{ for some } i \in S_0 \right) \\
&= 1 - \sum_h \int \mathrm{pr}\left(Q_i \leqslant \frac{\alpha}{N_0} \text{ for some } i \in S_0 | W^m = w, H^m = h \right) dM(w^m | h_j = 0)\pi(h) \\
&\geqslant 1 - \sum_h \int \sum_{i \in S_0(h)} \mathrm{pr}\left(Q_i \leqslant \frac{\alpha}{N_0} | W^m = w, H^m = h \right) dM(w^m | h_j = 0)\pi(h) \\
&= 1 - \sum_h \int \sum_{i \in S_0(h)} \mathrm{pr}\left(P_i \leqslant \frac{w_i \alpha}{N_0} | W^m = w, H^m = h \right) dM(w^m | h_j = 0)\pi(h) \\
&= 1 - \sum_h \int \sum_{i \in S_0(h)} \frac{w_i \alpha}{N_0} dM(w^m | h_j = 0)\pi(h) \\
&= 1 - \alpha \sum_h \int \frac{1}{N_0} \sum_{i \in S_0(h)} w_i dM(w^m | h_j = 0)\pi(h) \\
&= 1 - \alpha \sum_h \int \frac{N_0}{N_0} dM_0(w)\pi(h) = 1 - \alpha \sum_h \int dM_0(w)\pi(h) = 1 - \alpha.
\end{aligned}$$

□

Assume that A occurs and let $v = \min\{j : H_{(j)} = 0\}$. Then

$$Q_{(v)} > \frac{\alpha}{N_0} = \frac{\alpha}{\sum_{i \in S_0} W_i} \geqslant \frac{\alpha}{\sum_{i=v}^{m} W_{(i)}},$$

which implies that $S_0 \cap \mathscr{R}_0 = \varnothing$.

We now consider control of the false discovery exceedance at level c. Let $\mathscr{R}_0$ be the rejection region from either the weighted Bonferroni method or the weighted Holm method. Define $\mathscr{R}$ as follows. If $\#(\mathscr{R}_0) < (1-c)/c$ take $\mathscr{R} = \varnothing$; otherwise take $\mathscr{R} = \mathscr{R}_0 \cup A$, where $A \subset S$ is any set of hypotheses such that $A \cap \mathscr{R}_0 = \varnothing$ and $\#(A)/\{\#(A) + \#(\mathscr{R}_0)\} \leqslant c$.

THEOREM 3. *If $\mathscr{R}$ is constructed as defined above then* $\mathrm{pr}(\mathrm{FDP} > c) \leqslant \alpha$.

The proof is essentially the same as the proofs for the unweighted case in Genovese & Wasserman (2004) and the technical reports by Genovese and Wasserman and van der Laan et al.

There is freedom in choosing the extra rejections. Two alternatives are to choose the k hypotheses not in $\mathscr{R}_0$ with the smallest q-values or the smallest p-values, where $k \doteq \#(\mathscr{R}_0)c/(1-c)$. The former, which we use in simulations, will have somewhat higher power when the weights are well chosen and the latter will be more robust to misspecification of the weights. Based on the methods in the technical report by Genovese and Wasserman it is possible to construct versions with even higher power but we shall not pursue them here.

6. Power of the weighted procedures

Having established that our procedures control the false discovery rate or the false discovery exceedance, we next turn to the question of what effect weighting has on

power. To make weighting worthwhile, power should improve substantially when guessing is informative but should not drop too low when guessing is poor. The asymmetry between null and alternative makes this possible. With F stochastically smaller than the Uniform, assigning small weights to true alternatives can still allow the corresponding null hypotheses to be preferentially rejected, and similarly for large weights assigned to true nulls. This 'power arbitrage' does in fact appear to hold in practice. Indeed, if weighting is 'informative', in that the weights are positively associated with the null hypothesis being false, we would expect weighting to improve power over the corresponding unweighted procedure. In this section, we provide some theoretical support for this idea.

Let M_0 and M_1 be weight distributions as in § 2, with respective means μ_0 and μ_1. We assume that $\mu = (1-a)\mu_0 + a\mu_1 \equiv 1$, so that $0 < \mu_0 < 1/(1-a)$ and $0 < \mu_1 < 1/a$. We call $\mu_1 > 1$ the informative case and $\mu_1 < 1$ the misinformative case.

Define the Type I error rate and power as a function of the threshold:

$$I(t) = \mathrm{pr}(P \leqslant Wt | H = 0) = \mu_0 t,$$

$$H(t) = \mathrm{pr}(P \leqslant Wt | H = 1) = \int F(wt) dM(w|h = 1).$$

Note that the Type I error rate decreases with μ_0.

Let t^w and t^0 denote the asymptotic thresholds, defined from the population probability distributions, for the weighted and unweighted methods. The ratio of powers is then

$$\frac{H(t^w)}{F(t^0)} = \frac{t^w}{t_0}\left\{1 + (\mu_1 - 1)\frac{a\alpha}{1-(1-a)\alpha}\right\}. \tag{7}$$

To see this, note that, at the specified threshold,

$$\frac{t^w}{D(t^w)} = \alpha = \frac{t^0}{G(t^0)}.$$

Solving for $H(t^w)$ on the left-hand side yields $H(t^w)/t^w = 1/(a\alpha) - \mu_0(1-a)/a = 1/(a\alpha) + \mu_1 - 1/a$, by the relationship between μ_0 and μ_1. A similar calculation for $F(t^0)$ shows that $F(t^0)/t^0 = 1/(a\alpha) - (1-a)/a$.

Equation (7) is less than satisfying on its own because the thresholds t^0 and t^w depend on F and H as well. To investigate how the threshold changes with the weighting, we introduce a one-parameter family. For $0 \leqslant \lambda \leqslant 1$, let $W^\lambda = \lambda W + 1 - \lambda$. Then $EW^\lambda = 1$, $\bar{W}_m = 1$ implies that $\bar{W}^\lambda_m = 1$, and, for $j = 0, 1$,

$$\mu^\lambda_j = E(W^\lambda | H = j) = 1 + \lambda(\mu_j - 1).$$

Let t^λ be the asymptotic threshold defined by $t^\lambda / D^\lambda(t^\lambda) = \alpha$, where t^0 is the BH threshold and $D^0 = G$. Exploiting this equality as λ varies, we can define

$$R(t, \lambda) = \int \frac{F\{(\lambda w + 1 - \lambda)t\}}{t} dM(w|h = 1) - \lambda(\mu_1 - 1) - \frac{F(t^0)}{t^0}.$$

Then $R(t^\lambda, \lambda) \equiv 0$ for $0 \leqslant \lambda \leqslant 1$. Hence, we can differentiate $R(t^\lambda, \lambda)$ with respect to λ to obtain an expression for $dt^\lambda/d\lambda$. Computing the partial derivatives of R at $(t^0, 0)$ and applying the Implicit Function Theorem yields

$$\left.\frac{dt^\lambda}{d\lambda}\right|_{t^o,0} = (\mu_1 - 1)t_0 \frac{f(t_0) - t_0}{F(t_0) - t_0 f(t_0)},$$

where $F(t_0) - t_0 f(t_0) > 0$ by the strict concavity of F. To first order then, we have

$$\frac{t^\lambda}{t^0} = 1 + (\mu_1 - 1)t_0 \frac{f(t_0) - t_0}{F(t_0) - t_0 f(t_0)},$$

and plugging this into equation (7) gives to first order that

$$\begin{aligned} \frac{H(t^\lambda)}{F(t^0)} &= \left\{1 + (\mu_1 - 1)t_0 \frac{f(t_0) - t_0}{F(t_0) - t_0 f(t_0)}\right\}\left\{1 + (\mu_1 - 1)\frac{a\alpha}{1 - (1-a)\alpha}\right\} \\ &= \left\{1 + (\mu_1 - 1)t_0 \frac{f(t_0) - t_0}{F(t_0) - t_0 f(t_0)}\right\}\left\{1 + (\mu_1 - 1)t_0 \frac{1}{F(t_0)}\right\}. \end{aligned} \tag{8}$$

In the informative case, the only term that is possibly negative here is $f(t_0) - t_0$. By concavity, this is a measure of distance between the alternative F and the Uniform distribution. For fixed a and α, t^0 is determined by the intersection of F with a fixed line of slope $\{1 - (1-a)\alpha\}/(a\alpha)$. An F with mass concentrated near zero intersects that line where the density is small; an F close to the Uniform intersects the line at larger density. When $f(t_0)$ is very small, there is less room for improvement in power because most of the alternatives will have been rejected at t_0.

Consider the above expression for a parametric family of alternatives. If we assume that the test statistics are $N(\theta, 1)$ under the alternative, then the resulting p-values have density $f_\theta(t) = \exp\{-1/2\theta^2 + \theta\Phi^{-1}(1-t)\}$. For this model, $f(t_0) > t_0$ for any $\theta \leqslant 5$, at least. When

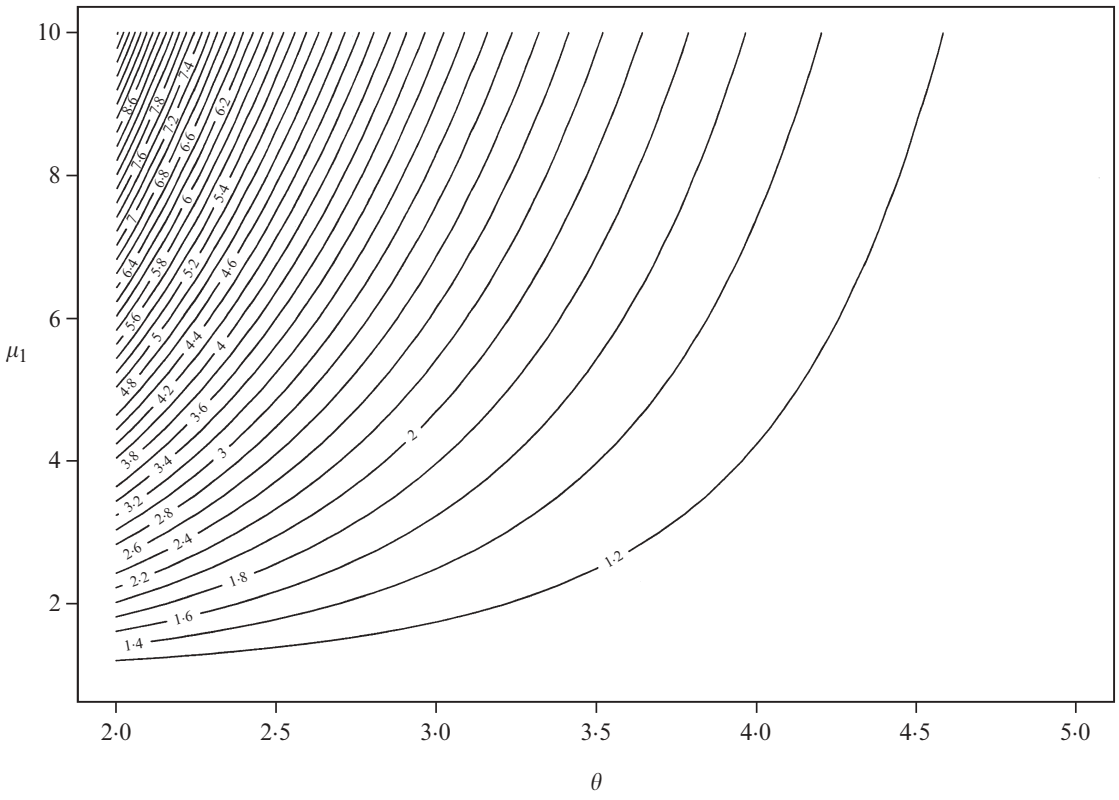

Fig. 1. Ratio of weighted to unweighted power in the $N(\theta, 1)$ family, for the procedures for controlling the false discovery rate from equation (8) as a function of μ_1 and θ, for $a = \alpha = 0{\cdot}05$.

$\theta > 5$, power is likely to be near one for either the wBH or the BH procedure. Thus, informative weighting is likely to improve power for large m. Figure 1 gives a representative contour plot of the power ratio as a function of θ and μ_1 in this family.

Furthermore, for small α, $f(t_0) > t_0$ so that informative weighting should improve power in this case as well. In particular, if $f(\alpha) > \alpha$, then, because $t_0 \leqslant \alpha$, $f(t_0) \geqslant f(\alpha) > \alpha \geqslant t_0$, and informative weighting improves power. In the $N(\theta, 1)$ family,

$$f_\theta(t) = \exp\{-1/(2\theta^2) + \theta\Phi^{-1}(1-t)\},$$

so that informative weighting improves power at least for all $\theta \leqslant \theta_\alpha$, where

$$\theta_\alpha = \Phi^{-1}(1-\alpha) + \sqrt{[\{\Phi^{-1}(1-\alpha)\}^2 - 2\log\alpha]}. \tag{9}$$

For example, $\theta_{0{\cdot}05} = 4{\cdot}59$ and $\theta_{0{\cdot}01} = 6{\cdot}15$; as Fig. 1 shows, these are conservative and in practice the boundary θ will be higher.

7. SIMULATION STUDIES

In this section, we present simulations and power calculations to evaluate the power of wBH under a variety of weighting regimes. We limit our attention to binary weighting schemes with $\gamma = a$ but allow for a wide range of η and r.

Figure 2 compares the power of wBH to the standard BH method and the BH 'Oracle', which takes a as known. We considered test statistics with $N(\theta, 1)$ alternatives for $\theta \in \{3, 4\}$

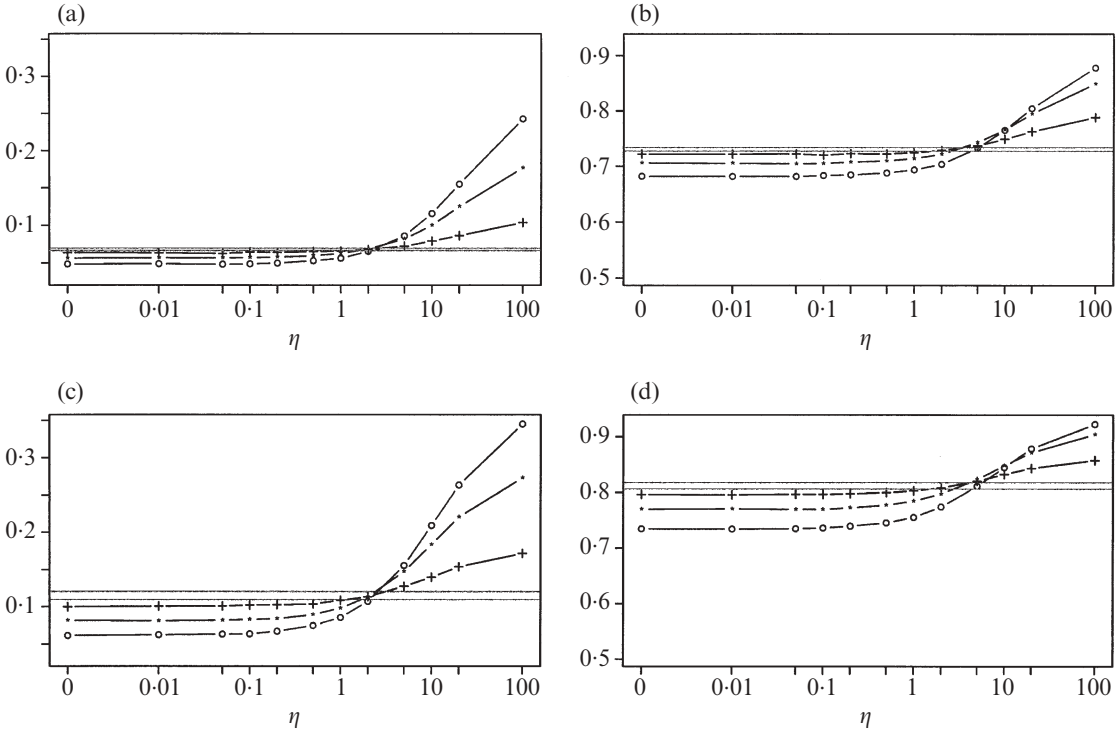

Fig. 2: Simulation study of power of the wBH procedure as a function of η and r. Power is depicted for four scenarios: (a) $\theta = 2$, $a = 0{\cdot}05$, (b) $\theta = 3{\cdot}5$, $a = 0{\cdot}05$, (c) $\theta = 2$, $a = 0{\cdot}1$ and (d) $\theta = 3{\cdot}5$, $a = 0{\cdot}1$. The BH and BH Oracle powers are given by the horizontal lines, with the latter always larger. The wBH method is evaluated for $\eta = 0$, 0·01, 0·05, 0·1, 0·2, 0·5, 1, 2, 5, 10, 20 and 100, and for three choices of weights: $r = 2$ (+), $r = 5$ (•) and $r = 10$ (◦). Vertical scales are held fixed across θ but not a.

and $a \in \{0{\cdot}01, 0{\cdot}05, 0{\cdot}1\}$; however, the results for $a = 0{\cdot}01$ are not shown, being similar in nature to those for $a = 0{\cdot}05$. We ran 10 000 iterations for each configuration, computing results for all methods on the same data. This amounts to 360 000 iterations for each of the BH results because they are not affected by η or r.

We briefly discuss the power results fully in § 8. The false discovery rate was controlled at the nominal level, within simulation error, for all cases. As expected from Theorem 1, wBH ensures that $\text{FDR} \leqslant (1-a)\mu_0\alpha$. Under informative guessing, therefore, it both improves power and reduces the false discovery rate below the nominal level. This suggests an adaptive method for gaining further power by estimating μ_0 from the data.

To investigate the power of the weighted Holm-based method for FDX control, we also conducted a simulation using the same settings as above. The results are given in Fig. 3.

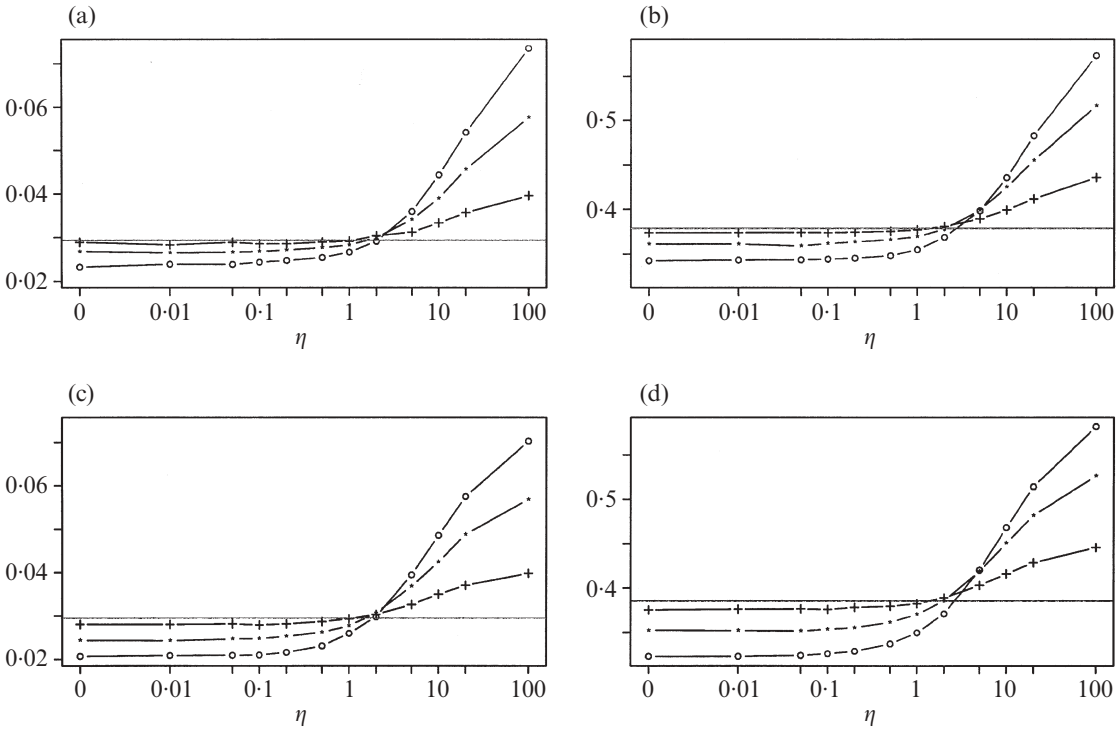

Fig. 3. Simulation study of power of the weighted FDX-controlled procedure as a function of η and r. Power is depicted for four scenarios: (a) $\theta = 2$, $a = 0{\cdot}05$, (b) $\theta = 3{\cdot}5$, $a = 0{\cdot}05$, (c) $\theta = 2$, $a = 0{\cdot}1$ and (d) $\theta = 3{\cdot}5$, $a = 0{\cdot}1$. The power of the weighted Holm procedure is given by the horizontal line. Power for the weighted procedure is evaluated for $\eta = 0$, 0·01, 0·05, 0·1, 0·2, 0·5, 1, 2, 5, 10, 20 and 100, and for three choices of weights: $r = 2$ (+), $r = 5$ (•) and $r = 10$ (◦). Vertical scales are held fixed across θ but not a.

8. Discussion

In cases where there is spatial structure among the tests, one approach is to focus on inference for significant regions. Perone Pacifico et al. (2004) show how to control the proportion of false regions in a random field context. In a personal communication, J. Taylor alters the null hypothesis to account for adjacency and devises a procedure that is better able to distinguish structure signal from scattered noise.

Our simulation results, see Figs 2 and 3, confirm that, while controlling the false discovery rate or the false discovery exceedance, weighting can lead to substantial gains

in power when the weights are well chosen ($\mu_1, \eta > 1$) whereas the potential loss in power is small even when the weights are poorly chosen ($\mu_1, \eta < 1$). Remarkably, the loss of power is not serious even if the weights are completely wrong ($\eta = 0$).

Much research has been done on estimating a to construct adaptive procedures with higher power than the BH procedure. However, the gain in power from such adaptive procedures is bounded above by the difference in heights of the two horizontal lines in Figs 2 and 3. Note that the potential gain in power from informative weighting is at least as large and often substantially larger.

Comparing Fig. 2 to Fig. 3, we see that control of the false discovery rate typically provides higher power than control of the false discovery exceedance, as expected given the stronger guarantee of the latter, but that the weighted false discovery exceedance procedures for controlling this are more robust to poorly chosen weights. The power below $\eta = 1$ drops by an essentially neglible amount. This suggests using a large value of r for control of the false discovery exceedance, whereas smaller r's seem warranted for control of the false discovery rate.

Other weighted multiple testing methods have been proposed. In contrast to Benjamini & Hochberg's (1997) weighted approach, our procedure aims to weight hypotheses highly if they are more likely to be nonnull, a priori. For a threshold T, the WBH procedure defines the weighted false discovery proportion as

$$\frac{\sum_{i=1}^{m} 1\{P_i \leqslant W_i T\}(1 - H_i)}{\sum_{i=1}^{m} 1\{P_i \leqslant W_i T\} + 1\{\text{all } P_i/W_i > t\}}.$$

Benjamini & Hochberg (1997) weight hypotheses based on the relative consequences of making a false discovery on the ith hypothesis. They define the weighted false discovery proportion as

$$\frac{\sum_{i=1}^{m} W_i 1\{P_i \leqslant T\}(1 - H_i)}{\sum_{i=1}^{m} W_i 1\{P_i \leqslant T\} + 1\{\text{all } P_i > t\}}.$$

With the former approach every false discovery is counted equally. The weights define varying thresholds for rejection, with heavily weighted hypotheses being rejected more easily. With the latter approach, heavily weighted hypotheses count more when we are assessing the false discovery rate, but all hypotheses are rejected or accepted with an equal threshold value. Clearly both of these approaches have merit, and which is preferable depends upon the context of the experiment.

In considering the use of prior information to improve testing, a Bayesian approach comes to mind. Indeed, the Bayesian method given in Genovese & Wasserman (2002b) can easily be extended to incorporate distinct priors for each hypothesis. Storey (2002) and Efron et al. (2001) have given Bayesian interpretations of the false discovery rate. It would be interesting to investigate the relationship between a weighted version of their procedures and a fully Bayesian approach.

ACKNOWLEDGEMENT

This work was partially supported by funding from the U.S. National Institutes of Health and the National Science Foundation. We thank the referees for their helpful comments.

Appendix

Proof of Theorem 2

For simplicity first assume that the W_i's are not constrained to have average 1, and are thus independent. Fix $b, \varepsilon > 0$. Then $C(t_* + b) < 1/\alpha$, so that $C(t_* + b) = 1/\alpha - \delta$ for some $\delta > 0$. We have, using Lemma 1, that, for every $t > t_* + b$,

$$\begin{aligned}\hat{C}_m(t) = \frac{\hat{D}_m(t)}{t\bar{W}_m} &\leqslant \frac{D(t) + \sup_u |\hat{D}_m(u) - D(u)|}{t\mu - t|\bar{W}_m - \mu|} \leqslant \frac{D(t) + \varepsilon}{t(\mu - \varepsilon)} \\ &= C(t)\left(\frac{\mu}{\mu - \varepsilon}\right) + \frac{\varepsilon}{t(\mu - \varepsilon)} \leqslant C(t_* + b)\left(\frac{\mu}{\mu - \varepsilon}\right) + \frac{\varepsilon}{(t_* + b)(\mu - \varepsilon)} \\ &= \left(\frac{1}{\alpha} - \delta\right)\left(\frac{\mu}{\mu - \varepsilon}\right) + \frac{\varepsilon}{(t_* + b)(\mu - \varepsilon)} < \frac{1}{\alpha}, \end{aligned} \tag{A1}$$

for large enough m. Equation (A1) follows for large m from the Gilvenko–Cantelli Theorem, which implies that $\sup_u |\hat{D}_m(u) - D(u)| \to 0$ almost surely, and the Strong Law of Large Numbers, which implies that $\bar{W}_m \to \mu$ almost surely. Hence, $T_m < t_* + b$. Combined with a similar argument applied to $t < t_* - b$, this implies that $|T_m - t_*| \leqslant b$ almost surely for all large m.

Now,

$$\text{FDP}(T_m) = \frac{m^{-1}\sum_i (1 - H_i) 1\{P_i/W_i \leqslant t\}}{m^{-1}\sum_i 1\{P_i/W_i \leqslant t\}} = \frac{\hat{V}_m(T_m)}{\hat{D}_m(T_m)}.$$

Then, for all large m,

$$\begin{aligned}|\hat{D}_m(T_m) - D(t_*)| &= |\hat{D}_m(T_m) - D(T_m) + D(T_m) - D(t_*)| \\ &\leqslant \sup_u |\hat{D}_m(u) - D(u)| + |D(T_m) - D(t_*)| \to 0\end{aligned}$$

by the Gilvenko–Cantelli and Continuous Mapping Theorems. By a similar argument, $\hat{V}_m(T_m) - V(t_*) = o(1)$ almost surely, where

$$V(t) = E(1 - H_i) 1\{P_i/W_i \leqslant t\} = (1 - a)t\mu_0 \leqslant t\mu.$$

If the W_i's are constrained to have average 1, we proceed as follows. Write $W_i = U_i/\bar{U}_m$ for independent and identically distributed variables $U_1, \ldots, U_m$. Let $\tilde{W}_i = U_i/EU_1$. It follows that

$$\begin{aligned}\hat{D}_m(t) &= \frac{1}{m}\sum_i 1\left\{\frac{P_i}{W_i} \leqslant t\right\} = \frac{1}{m}\sum_i 1\left\{\frac{P_i}{\tilde{W}_i} \leqslant t\frac{EU_1}{EU_1 + (\bar{U}_m - EU_1)}\right\} \\ &\leqslant \frac{1}{m}\sum_i 1\left\{\frac{P_i}{\tilde{W}_i} \leqslant t\frac{EU_1}{EU_1 - \varepsilon}\right\} \leqslant \text{pr}\left(\frac{P_i}{\tilde{W}_i} \leqslant t\frac{EU_1}{EU_1 - \varepsilon}\right) + \varepsilon,\end{aligned}$$

for large enough m, uniformly in t. Similarly,

$$\hat{D}_m(t) \geqslant \frac{1}{m}\sum_i 1\left\{\frac{P_i}{\tilde{W}_i} \leqslant t\frac{EU_1}{EU_1 + \varepsilon}\right\} \geqslant \text{pr}\left(\frac{P_i}{\tilde{W}_i} \leqslant t\frac{EU_1}{EU_1 + \varepsilon}\right) - \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that $\sup_u |\hat{D}_m(t) - \tilde{D}_m(t)| \to 0$ almost surely, where $\tilde{D}(t)$ is the empirical probability distribution of $\text{pr}(P_i/\tilde{W}_i \leqslant t)$ and $\tilde{D}$. The remainder of the proof is the same.

Thus, $|\text{FDP}(T_m) - \text{FDP}(t_*)| \to 0$ almost surely, and, because this is bounded, dominated convergence yields the second claim. The third claim follows immediately.

REFERENCES

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

BENJAMINI, Y. & HOCHBERG, Y. (1997). Multiple hypothesis testing with weights. *Scand. J. Statist.* **24**, 407–18.

BENJAMINI, Y. & YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.

BENJAMINI, Y., KRIEGER, A. & YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507.

EFRON, B., STOREY, J., TUSHER, V. G. & TIBSHIRANI, R. (2001). Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Assoc.* **96**, 1151–60.

GENOVESE, C. R. & WASSERMAN, L. (2002a). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc.* B **64**, 499–518.

GENOVESE, C. R. & WASSERMAN, L. (2002b). Bayesian and frequentist multiple testing. In *Bayesian Statistics VI*, Ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, pp. 145–62. Oxford: Clarendon Press.

GENOVESE, C. R. & WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035–61.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

KORN, E. L., TROENDLE, J. F., MCSHANE, L. M. & SIMON, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *J. Statist. Plan. Infer.* **124**, 379–98.

PERONE PACIFICO, M., GENOVESE, C., VERDINELLI, I. & WASSERMAN, L. (2004). False discovery control for random fields. *J. Am. Statist. Assoc.* **99**, 1002–14.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc.* B **64**, 479–98.

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

[*Received November* 2004. *Revised January* 2006]