Exceedance Control of the False Discovery Proportion

Christopher Genovese[1] and Larry Wasserman[2]
Carnegie Mellon University
July 10, 2004

Multiple testing methods to control the False Discovery Rate (FDR), the expected proportion of falsely rejected null hypotheses among all rejections) have received much attention. It can be valuable instead to control not the mean of this false discovery proportion (FDP) but the probability that the FDP exceeds a specified bound. In this paper, we construct a general class of methods for exceedance control of FDP based on inverting tests of uniformity. The method also produces a confidence envelope for the FDP as a function of rejection threshold. We discuss how to select a procedure with good power.

KEYWORDS: Multiple Testing, p-values, False Discovery Rate.

1 Introduction

Multiple testing procedures to control the False Discovery Rate, first introduced by Benjamini and Hochberg (1995), have recently received much attention: see Benjamini and Yekutieli (2001), Efron, Tibshirani and Storey (2002), Finner and Roters (2002), Sarkar (2002), Storey (2002, 2003), and Storey, Taylor and Siegmund (2003). These methods can attain high power even when testing thousands or millions of hypotheses. They are ideal for large scale multiple testing problems that occur in bioinformatics, imaging and many other areas.

The False Discovery Rate (FDR) is the expected value of what we call the False Discovery Proportion (FDP), namely the proportion of falsely rejected null hypotheses among all rejected null hypotheses. In some cases, it can be useful to control, not the expected FDP, but the probability that the FDP exceeds a specified bound. We call this *exceedance control* of the FDP.

[1] Research supported by NSF Grant SES 9866147.

[2] Research supported by NIH Grant R01-CA54852-07 and NSF Grant DMS-98-03433 and NSF Grant DMS-0104016.

Genovese and Wasserman (GW, 2004) introduced a method for exceedance control based on a confidence envelope for the unobserved FDP. Perone Pacifico, Genovese, Verdinelli and Wasserman (PGVW, 2003) extended this approach to random fields. And van der Laan, Dudoit and Pollard (VDP, 2004) proposed a method based on augmenting familywise tests.

In this paper, we extend the approach of GW and PGVW, leading to a class of procedures that can achieve higher power and handle arbitrary dependence among tests. In particular, we find that using a test statistic based on the minimum p-value has suboptimal power. We also examine the relationship between this approach and the VDP method.

2 Background

Let $X_1, \ldots, X_n$ be random vectors drawn IID from a distribution $\mathbb{P}$. We consider m hypotheses of the form

$$H_{0j} : \ \mathbb{P} \in \mathcal{M}_j \quad \text{versus} \quad H_{1j} : \ \mathbb{P} \notin \mathcal{M}_j \qquad j = 1, \ldots, m, \tag{1}$$

for sets of probability distributions $\mathcal{M}_1, \ldots, \mathcal{M}_m$. (The hypothesis testing testing notation from VDP is rather elegant, and we will use a similar notation.) For the typical cases we have in mind, $m >> n$. A common case is when each vector $X_i = (X_{i1}, \ldots, X_{im})$ comprises m measurements on subject i. For example, in microarray studies, X_{ij} might be the gene expression level of gene j for subject i; in brain imaging studies, X_{ij} might be a statistic computed at brain location j for subject i; and in astronomical imaging, X_{ij} might be a photon count at sky location j for session i. In each of these examples, the $\mathcal{M}_j$s might be defined as $\mathcal{M}_j = \{\mathbb{P} : \ \mathbb{E}_{\mathbb{P}}(X_{ij}) = \mu_j\}$ for some constant μ_j.

Define hypothesis indicator variables $H^m = (H_1, \ldots, H_m)$ such that $H_j = 1\{\mathbb{P} \notin \mathcal{M}_j\}$. Let $S = \{1, \ldots, m\}$ and let

$$S_0 \equiv S_0(\mathbb{P}) = \{j : \ H_j = 0\} \tag{2}$$

be the set of true nulls. For each $j \in S$, let $Z_j = Z_j(X_1, \ldots, X_n)$ be a test statistic for the null hypothesis H_{0j}. Let $P^m = (P_1, \ldots, P_m)$ denote the corresponding p-values. We assume that if $H_j = 0$ then $P_j \sim \text{Unif}(0,1)$. Denote the ordered p-values by $P_{(1)} < \cdots < P_{(m)}$, and define $P_{(0)} \equiv 0$. For any $W \subset S$, define $P_W = (P_i : \; i \in W)$.

We call any $R = R(P_1, \ldots, P_m) \subset S$ a *rejection region* and say that R controls familywise error rate at level α if

$$\mathbb{P}\{\#(R \cap S_0(\mathbb{P})) > 0\} \leq \alpha,$$

where $\#(B)$ denotes the number of points in a set B. More generally, say that R controls the k-familywise error rate at level α if

$$\mathbb{P}\{\#(R \cap S_0(\mathbb{P})) > k\} \leq \alpha.$$

We define the false discovery proportion (FDP) of a rejection set R by

$$\Gamma(R) \equiv \text{FDP} = \frac{\text{false rejections}}{\text{rejections}} = \frac{\sum_{j=1}^{m}(1 - H_j)1\{R \ni j\}}{\sum_{j=1}^{m} 1\{R \ni j\}} \tag{3}$$

where the ratio is defined to be zero if the denominator is zero. The false discovery rate FDR is defined by $\text{FDR} = \mathbb{E}(\text{FDP})$.

Our goal in this paper is to find a rejection region $R = R(P_1, \ldots, P_m)$ such that

$$\mathbb{P}\{\Gamma(R) > c\} \leq \alpha \tag{4}$$

for given c and α. We call such an R a (c, α) *rejection region.* Typically, $R = \{j \in S : \; P_j \leq T\}$ for some random threshold $T = T(P_1, \ldots, P_m)$, in which case we may write $\Gamma(T)$ for the FDP. A $1 - \alpha$ *confidence envelope* for FDP is a random function $\overline{\Gamma}(C) = \overline{\Gamma}(C; P_1, \ldots, P_m)$ such that

$$\mathbb{P}\Big\{\overline{\Gamma}(C) \geq \Gamma(C), \text{ for all } C\Big\} \geq 1 - \alpha. \tag{5}$$

For rejection regions based on a fixed p-value threshold t, it is convenient to write $\overline{\Gamma}(t)$ and $\Gamma(t)$. Specifying the function $t \mapsto \overline{\Gamma}(t)$ is sufficient to determine the entire envelope for rejection regions of the form $\{j \in S : \; P_j \leq T\}$.

3 Exceedance Control

In this section, we describe two approaches to controlling FDP exceedance and then we show that the two are related. The first, called *inversion*, was first described in GW. This produces a confidence envelope for FDP by inverting a set of uniformity tests. The second, called *augmentation*, was described in VDP. This produces a rejection region controlling FDP exceedance by expanding the rejection region from a familywise test.

The inversion method involves the following steps:

1. For every $W \subset S$, test at level α the hypothesis that $P_W = (P_i : i \in W)$ is a sample from a Uniform$(0,1)$ distribution:
$$H_0 : W \subset S_0 \quad \text{versus} \quad H_1 : W \not\subset S_0. \tag{6}$$
Formally, let $\Psi = \{\psi_W : W \subset S\}$ be a set of non-randomized tests such that $\mathbb{P}\big\{\psi_W(U_1, \ldots, U_{\#(W)}) = 1\big\} \leq \alpha$ whenever $U_1, \ldots, U_{\#(W)} \sim$ Unif$(0,1)$.

2. Let $\mathcal{U}$ denote the collection of all subsets W not rejected in the previous step:
$$\mathcal{U} = \{W : \psi_W(P_W) = 0\}. \tag{7}$$

3. Define
$$\overline{\Gamma}(C) = \begin{cases} \max\limits_{B \in \mathcal{U}} \dfrac{\#(B \cap C)}{\#(C)} & \text{if } C \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

4. Choose $R = R(P_1, \ldots, P_m)$ such that $\overline{\Gamma}(R) \leq c$. (Typically, R is of the form $R = \{j : P_j \leq T\}$ where the *confidence threshold* $T = \sup\{t : \overline{\Gamma}(t) \leq c\}$.)

It follows that $\overline{\Gamma}$ is a $1 - \alpha$ confidence envelope for FDP and R is a (c, α) rejection set.

If $\mathcal{U}$ is closed under unions, then
$$\overline{\Gamma}(C) = \frac{\#(U \cap C)}{\#(C)} \tag{9}$$
where $U = \cup\{V : V \in \mathcal{U}\}$. Moreover, U is a *confidence superset* for S_0 in the sense that
$$\mathbb{P}\{S_0 \subset U\} \geq 1 - \alpha. \tag{10}$$

One can also use exceedance control for FDR:

LEMMA 3.1 (PGVW). *Let $c \in (0, \alpha)$. If $\mathbb{P}\{\Gamma(T) > c\} \leq \beta$ and $\beta = (\alpha - c)/(1 - c)$ then $\mathbb{E}(\Gamma(T)) \leq \alpha$.*

REMARK 3.1. One can choose any level α uniformity test in (6), but the choice of test does impact the results. Starting in the next section, we discuss this issue in detail.

The augmentation approach of VDP is described in the next theorem.

THEOREM 3.1. *(VDP) Suppose that R_0 is a rejection region that controls familywise error at level α. If $R_0 = \emptyset$ take $R = \emptyset$. Otherwise, let A be a set with $A \cap R = \emptyset$ and set $R = R_0 \cup A$. Then, $\mathbb{P}\{\Gamma(R) > c\} \leq \alpha$, where $c = \#(A)/(\#(A) + \#(R_0))$.*

The same logic easily gives a confidence envelope, as follows.

THEOREM 3.2. *Suppose that $R_0 = \{j : \ P_j \leq Q\}$ for some Q and that R_0 controls familywise error at level α. Define*

$$\overline{\Gamma}(C) = \begin{cases} \dfrac{\#(C - R_0)}{\#(C)} & \textit{if } C \neq \emptyset, \\ 0 & \textit{otherwise.} \end{cases} \tag{11}$$

Then, $\overline{\Gamma}$ is a $1 - \alpha$ confidence envelope for FDP.

The following result generalizes the augmentation method. The proof is straightforward.

THEOREM 3.3. *Suppose that R_k controls k-familywise error at level α. Let A be a set with $A \cap R_k = \emptyset$, and set $R = R_k \cup A$. Then, $\mathbb{P}\{\Gamma(R) > c\} \leq \alpha$ where $c = (\#(A) + k)/(\#(A) + \#(R_k))$.*

The relationship between inversion and augmentation is explained in the following two results.

THEOREM 3.4. *Let $\overline{\Gamma}_{\text{aug}}$ be the $1 - \alpha$ confidence envelope from (11) and let R_{aug} be such that $\overline{\Gamma}_{\text{aug}}(R_{\text{aug}}) \leq c$. Let $\overline{\Gamma}_{\text{inv}}$ be the $1 - \alpha$ confidence envelope from (8) with the test in (6) is defined as follows: $\psi_W(P_W) = 1$ if $R_0 \cap W \neq \emptyset$. Then $\overline{\Gamma}_{\text{aug}} = \overline{\Gamma}_{\text{inv}}$ and in particular, $\overline{\Gamma}_{\text{inv}}(R_{\text{aug}}) \leq c$.*

There is also a partial converse to the theorem above whose proof is straightforward.

THEOREM 3.5. *Let Ψ be a class of uniformity tests and let $\overline{\Gamma}_{\rm inv}$ be the resulting $1-\alpha$ confidence envelope and let $R_{\rm inv}$ be such that $\overline{\Gamma}_{\rm inv}(R_{\rm inv}) \leq c$. Let $\mathcal{U}$ be the unrejected sets as defined in (7) and suppose that $\mathcal{U}$ is closed under unions. Let $U = \cup_{V \in \mathcal{U}} V$. Then $R_0 = U^c$ controls familywise error at level α. Let $\Gamma_{\rm aug}$ be the resulting augmentation envelope defined by (11). Then, $\overline{\Gamma}_{\rm inv} = \overline{\Gamma}_{\rm aug}$. Moreover, $R_{\rm inv} = R_0 \cup A$ where $A = R_{\rm inv} - R_0$, and hence $R_{\rm inv}$ is an augmentation rejection set.*

Hence, we see that the two methods lead to the same rejection regions. That is, the rejection region from any given augmentation procedure, is the rejection region of some inversion procedure. Conversely, under suitable conditions on the tests, the rejection set from any given inversion procedure, is the rejection region of some augmentation procedure. The exception is when $\mathcal{U}$ is not closed under unions. In that case, the inversion method produces a valid rejection region but it does not correspond to an augmentation of a familywise test. In the next section we discuss specific choices of Ψ, some of which have this property.

4 The $P_{(k)}$ Tests

In choosing the tests Ψ for (6), we have two goals:

1. (Power). The envelope $\overline{\Gamma}$ should be close to Γ and thus result in rejection regions with high power.
2. (Computational Tractability). The envelope $\overline{\Gamma}$ should be easy to compute.

Until Section 8, we assume the p-values are independent.

Now we discuss how to choose the class of tests Ψ. We expect significant results to manifest themselves as small p-values. This suggests using a test that is sensitive to departures from uniformity in the left tail. An obvious choice is to use the minimum order statistic $P_{(1)}$ as in PGVW and VDP.

A natural generalization is to use the k^{th} order statistic $P_{(k)}$. We will see that taking $k > 1$ can yield procedures with mich higher power than $k = 1$. Traditional uniformity tests, like the Kolmogorov-Smirnov test, do not fare well. The Kolmogorov-Smnirov test looks for deviations from uniformity equally though all the p-values. The following theorem gives the envelope $\overline{\Gamma}_k$ that results from using $P_{(k)}$ to test (6).

Let $\phi_{k,r}(v_1, \ldots, v_r)$ be the k^{th} smallest value of $v_1, \ldots, v_r$ or 1 if $r < k$. Let $\psi_W(P_W) = 1$ if

$$\phi_{k,\#(W)}(P_W) < B^{-1}_{k,\#(W)}(\alpha) \tag{12}$$

where $B_{a,b}$ is the CDF of a Beta(a, b) random variable. We call $\Phi = \{\phi_W : W \subset S\}$ the $P_{(k)}$ *test.*

THEOREM 4.1. *(The $P_{(k)}$ test.) The inversion-based confidence envelope $\overline{\Gamma}_k$ is as follows. Define*

$$J_k = \min\left\{ j : \; P_{(j)} \geq B^{-1}_{k,m-j+1}(\alpha) \right\}, \tag{13}$$

or $J_k = m + 1$ if no such j exists. Then,

$$\overline{\Gamma}_k(C) = \begin{cases} \dfrac{\#(\{\pi(k), \ldots, \pi(J_k)\} \cap C)}{\#(C)} & \text{if } J_k \leq m \\ 1 & \text{otherwise} \end{cases} \tag{14}$$

where $\pi(j)$ is defined by

$$P_{\pi(j)} = P_{(j)}. \tag{15}$$

In particular, with $t_k = P_{(J_k)}$,

$$\overline{\Gamma}_k(t) = \begin{cases} 1 & t \leq P_{(k-1)} \\ \frac{k-1}{m\widehat{G}(t)} & P_{(k-1)} < t \leq t_k \\ 1 - \frac{J_k - k + 1}{m\widehat{G}(t)} & t > t_k \end{cases} \tag{16}$$

where $\widehat{G}_m$ is the empirical CDF *of the p-values. A (c, α) rejection region is $R_k = \{j : \; P_j \leq T_k\}$ where $T_k = \sup\{t : \; \overline{\Gamma}_k(t) \leq c\}$ and $T_k = 0$ if no such t exists.*

Lemma 4.1. *When $k > 1$, the collection $\mathcal{U}$ derived from Ψ need not be closed under unions.*

Remark 4.1. The method in PGVW corresponds to taking $k = 1$. The augmentation method in VDP with R_0 based on their step-down procedure corresponds to the $k = 1$ method, except that the Beta quantiles are replaced by bootstrap estimated quantiles.

Theorem 4.2. *For $k > 1$, the rejection region $R_k = \{j : \ P_j \leq P_{(J_k)}\}$ controls $k-1$ familywise error rate at level α and the set $R = \{P_j : \ P_j \leq T_k\}$ corresponds to augmenting R_k with $A = \{j : \ P_{(J_k)} < P_j \leq T_k\}$.*

Figure 1 shows some plots of Γ_k for selected values of k. Notice that the curves are anti-unimodal (except $k = 1$) and that as k increases, the minimum gets larger but moves to the right. This means that T_k moves to the right (leading to higher power) unless the minimum is above c in which case $T_k = 0$. Thus, the optimal k involves a delicate tradeoff. We return to this in the next section.

5 Power and Optimality

The $k = 1$ test corresponds to using the maximum test statistic over each subset. This is common practice; see for example GW, PGVW, and VDP. The literature on testing random fields also uses the supremum of a test process which is the continuous analogue of $k = 1$.

The following heuristic suggests that $k = 1$ might be sub-optimal. Let $R = \{j : \ P_j \leq T_m\}$ be the rejection region using a $k = 1$ procedure. Now make 1000 copies of each p-value. The resulting increase in m causes the Beta quantiles (or whatever quantiles are used by the procedure) to shrink towards 0, and hence the proportion of rejections decreases. In contrast, it is easy to see from the results in Genovese and Wasserman (2002) that the Benjamimi-Hochberg procedure rejects exactly the same proportion of hypotheses.

In this section, we show that the $k = 1$ test – indeed, any fixed k – is, in general, sub-optimal. We begin by introducing a specific model for the p-values that permits some simple analysis. There are at least two, interesting asymptotic regimes. Recall that the p-values are based on test statistics depending on a sample of size n. In the first regime, we let n stay fixed but let m grow. We believe this is the most interesting and relevant regime for large scale testing problems. In the second regime, we also let n get large. Assuming the test statistics correspond to consistent tests, and that n grows sufficiently quickly with respect to m, this will force the p-values of the non-null hypothesis to approach 0. Eventually this will cause all the alternative p-values to become smaller than all the null p-values. This is the asymptotic regime studied in (VDP). Our emphasis is on the first case.

We consider the following model for the p-values. Let $H_1, \ldots, H_m \sim \text{Bernoulli}(a)$, $P_j \mid H_j = 0 \sim \text{Uniform}(0,1)$ and $P_j \mid H_j = 1 \sim F_j$ for some F_j. Further, assume that the F_j's are randomly drawn distributions from an arbitrary probability measure over the set of all CDFs. It follows that the marginal distribution of P_j is

$$G = (1-a)U + aF \tag{17}$$

where $U(p) = p$ and $F(p) = \mathbb{E}(F_j(p))$.

Since our goal is only to establish that any fixed k need to be optimal, it suffices to specialize the model to permit simpler analysis. We take $F = F_\beta(t)$, the CDF of a Uniform distribution on $[0, 1/\beta]$ with $\beta > 1$. The fixed n regime corresponds to keeping β fixed. The large n regime is obtained by letting $\beta \to \infty$. We show that fixed k procedures are sub-optimal in this class of problems. First we establish the optimal threshold.

Theorem 5.1. *Suppose that $(1-a)/\xi < c$ where*

$$\xi = (1-a) + a\beta > 1. \tag{18}$$

Let

$$T_* = \begin{cases} t_m & \text{if } (1-a)/\xi < c \leq 1-a \\ 1 & \text{if } c > 1-a \end{cases} \tag{19}$$

where

$$t_m = \left(\frac{a}{1-a}\right)\left(\frac{c - \frac{\sigma z_\alpha}{\sqrt{m}}}{1 - c + \frac{\sigma z_\alpha}{\sqrt{m}}}\right) \geq t_0, \tag{20}$$

$$t_0 = \left(\frac{a}{1-a}\right)\left(\frac{c}{1-c}\right), \tag{21}$$

and $\sigma > 0$ is described in the proof. Then:

1. *$R_* = \{j : \; P_i \leq T_*\}$ is an asymptotic (c, α) rejection region, that is,*

$$\limsup_{m\to\infty} \mathbb{P}\{\Gamma(R) > c\} \leq \alpha.$$

2. *R_* is optimal in the following sense: if $R = \{j : \; P_i \leq T_m\}$ is an asymptotic (c, α) rejection region then*

$$\mathbb{P}\{R \subset R_*\} \geq 1 - \alpha.$$

The following result shows that T_k is far from the optimal threshold T_*.

Theorem 5.2. *For any fixed k, the $P_{(k)}$ threshold $T_k = o_P(1)$ and $T_*/T_k \xrightarrow{p} \infty$.*

6 Combining $P_{(k)}$ tests

In many cases, the $P_{(1)}$ test can yield poor power. Using a larger k can in some cases improve power but there is a risk that the rejection region might be empty. In particular, $\min_t \overline{\Gamma}_k(t) = (k-1)/J_k$, and if $(k-1)/J_k > c$ then $R_k = \emptyset$.

As shown in the previous section, using any fixed k is suboptimal. Ideally, one could estimate the optimal k and use the corresponding $P_{(\widehat{k})}$, but this has two basic problems. When $\widehat{k}$ is larger than the optimal k, the rejection region can be trivial, leading to rather fragile performance. Also, the stochastic dependence between $\widehat{k}$ and $\overline{\Gamma}$ complicates analysis of the coverage properties.

A reasonable compromise is to combine $P_{(k)}$ envelopes over a range of ks. This ensures reasonable power over at least one k while maintaining a nontrivial rejection region.

The algorithm is as follows. Let $Q_m \subset \{1, \ldots, m\}$ be a set of integers and let q_m be the number of elements in Q_m. For each $k \in Q_m$, let $\overline{\Gamma}_k$ be the confidence envelope for FDP based on the $P_{(k)}$ test at level α/q_m. Let

$$\overline{\Gamma} = \min_{k \in Q_m} \overline{\Gamma}_k. \tag{22}$$

It follows that $\overline{\Gamma}$ is a valid confidence envelope for Γ and $R = \{j : \ P_j \leq T\}$ is a (c, α) rejection set, where $T = \sup\{t : \ \overline{\Gamma}(t) \leq c\}$.

In our examples, we take $Q_m = \{1, \ldots, \lfloor \widehat{a} c m \rfloor\}$, where $\widehat{a}$ is an estimate of the fraction of alternatives, such as $\widehat{a} = 2(\widehat{G}(1/2) - 1/2)$ which converges to $\underline{a} = 2(G(1/2) - 1/2) \leq a$; see Storey (2002). Technically, we should adjust the envelope to account for the randomness of $\widehat{a}$ but since $\widehat{a} - \underline{a} = O_P(m^{-1/2})$, this has a neglibible affect on the coverage as confirmed by our simulation studies in the next section.

Remark 6.1. We also considered an approach based on data splitting, where the optimal k is estimated from the training set and $P_{(\widehat{k})}$ is applied to the test set. But the performance of this method was poor; variability in $\widehat{k}$ made this much more fragile than even using $P_{(\widehat{k})}$ for the entire sample.

7 Simulation Studies

In this section, we report some simulation studies that show that the fixed k envelopes do not have uniformly reliable performance but that the combined procedure does.

The simulations illustrate a simple case of the mixture model (17) where the alternative test statistics are drawn IID from a Normal$(\theta, 1)$ and the nulls are drawn IID from a Normal$(0, 1)$. Using 1000 iterations in each configuration, we compute the mean FDP and power (the proportion of true alternatives rejected). In all cases, the coverage was controlled as predicted by the theory, that is, FDP ≤ 0.2 with probability at least 0.95.

Results are given in Table 1 for the combined procedure, $P_{(1)}$, and $P_{(10)}$. We computed other $P_{(k)}$ tests as well and the results are simialr.

8 Dependence

So far we have assumed that the p-values are independent. Extending the method to handle dependence is straightforward. We continue to assume that the marginal distribution of each P_j is Unif(0,1) under the null, but, we allow the joint distribution to be arbitrary.

THEOREM 8.1. *Replace J_k in equation (13) with*

$$J_k = \min\left\{ j : \; P_{(j)} \geq \frac{k\alpha}{m-j} \right\}. \tag{23}$$

Then the $P_{(k)}$ procedure (and its extensions in Section 6) are valid for arbitrary dependence among the p-values.

The above result follows from the earlier results together with the fact that if $Y_1, \ldots, Y_q$ are such that $Y_j \sim \text{Unif}(0,1)$, then $\mathbb{P}\big\{Y_{(k)} \leq c\big\} \leq cq/k$. To see this, let N be the number of Y_js less than c. Then,

$$\mathbb{P}\big\{Y_{(k)} \leq c\big\} = \mathbb{P}\{N \geq k\} \leq \frac{\mathbb{E}N}{k} = \frac{\sum_{j=1}^q \mathbb{P}\{Y_j \leq c\}}{k} = \frac{cq}{k}.$$

This upper bound is achieved by the following distribution. Draw $Y_j \sim \text{Unif}((j-1)k/q, jk/q)$ for $j = 1, \ldots, q/k$. Now, create $k-1$ copies of each Y_j and call these $Y_{q/k+1}, \ldots, Y_q$.

A different approach to dependence is suggested by VDP. Then estimate the quantiles of $P_{(k)}$ either by bootstrapping or by estimating the covariance matrix of the underlying statistics $Z_1, \ldots, Z_m$. This is less conservative than using the bound $\mathbb{P}\big\{Y_{(k)} \leq c\big\} \leq cq/k$. But unless n is very large compared to m, such an approach will be unlikely to succeed. Indeed, one needs at least $O(m^2)$ observations to estimate the covariance matrix.

9 FNP

Our methods can also be used to bound the false nondiscovery proportion (FNP). The FNP of a rejection region R is defined by

$$\Lambda(R) = \frac{\sum_{j=1}^m H_j 1\{R \not\ni j\}}{\sum_{j=1}^m 1\{R \not\ni j\}} \tag{24}$$

where the ratio is defined to be zero if the denominator is zero. Let $S_1 = S_0^c$. For each $A \subset S$ we test

$$H_0 : A \subset S_1 \quad \text{versus} \quad H_1 : A \not\subset S_1.$$

Define

$$\overline{\Lambda}(B) = \sup_{A \in \mathcal{U}} \frac{\#(A \cap B)}{\#(B)} \tag{25}$$

where $\mathcal{U}$ is the set of non-rejected A. Then $\overline{\Lambda}$ is a $1-\alpha$ confidence envelope for Λ.

In general, it is not possible to find a non-trivial confidence envelope that is valid for all CDFs F for the same reason that one cannot bound the power of a test unless the alternative is forced to be some distance from the null. Thus, let V be an invertible function such that $1 \geq V(t) \geq t$ for $0 \leq t \leq 1$ and define

$$\mathcal{F} = \{\text{CDF}s\ F :\ F \geq V\}.$$

Our goal is to find an envelope valid over all $\mathcal{F}$. For example, suppose that the test statistics are N(0,1) under the null and $N(\theta, 1)$ under the alternative. If we consider all alternatives such that $\theta > \theta_0$ for some fixed $\theta_0 > 0$, then the CDF F of the p-value distribution under every alternative, satisfies

$$F(t) \geq V(t) \equiv S(S^{-1}(t) - \theta_0)$$

where $S(t) = 1 - \Phi(t)$ and $\Phi(t)$ is the CDF of a standard Normal.

We will develop here the analogue of the $k = 1$ procedure. The extension to other k is similar to the FDP case. We begin by assuming independent p-values.

Theorem 9.1. *Let*

$$c(\alpha, \ell) = V^{-1}\Big((1-\alpha)^{1/\ell}\Big). \tag{26}$$

Define the test

$$\psi(A) = \begin{cases} 1 & \text{if } \max_{i \in A} P_i \ > \ c(\alpha, \#(A)) \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

Suppose we reject $H_0 : A \subset S_1$ when $\psi(A) = 1$. Then,

$$\sup_{F \in \mathcal{F}} \mathbb{P}_F(\psi(S_1) = 1) \leq \alpha. \tag{28}$$

Let $\overline{\Lambda}$ be the envelope from this test.

COROLLARY 9.1. *Using the above test, $\overline{\Lambda}$ is a valid $1 - \alpha$ confidence envelope for Λ.*

THEOREM 9.2. *Using the above test,*

$$\overline{\Lambda}(t) = \begin{cases} \frac{1-\widehat{G}(s)}{1-\widehat{G}(t)} & t < s \\ 1 & t \geq s \end{cases} \tag{29}$$

where $s = P_{(r)}$ and

$$r = \min\{j : \ P_{(j)} \leq c(\alpha, m - j + 1)\}. \tag{30}$$

THEOREM 9.3. *The above results are valid in the dependent case if we replace $c(\alpha, \ell)$ with*

$$c(\alpha, \ell) = V^{-1}(1 - \alpha). \tag{31}$$

10 Discussion

This paper, together with GW, PGVW and VDP, show that there is a rich class of methods for controlling FDP exceedance. This expands the set of tools available for false discovery control in multiple testing. While FDP exceedance methods can be more conservative than FDR-controlling procedures, they give a stronger guarantee and can be tuned to achieve a desired level of confidence.

One of the key challenges for false discovery control is handling dependence among the tests while maintaining reasonable power. We have presented variants of our method designed for the completely independent case and for arbitrary dependence. If more information is available about the nature of the dependence, it is possible that power can be improved, though to what degree is an open question. One approach might be to estimate the covariance and construct the quantiles for the uniformity tests based on the estimated covariance. VDP did this in the $k = 1$ case, assuming no prior information on the covariance. Such a method will typically require $n >> m$, which may not be realized in practice. We suspect that this constraint can be loosened with structural assumptions, such as local covariance in spatial problems.

It is an open question which of the available methods is optimal in which situations, but one result that seems evident is that using the maximum test statistic (e.g., $P_{(1)}$) is suboptimal in general.

11 Proofs

In this section, we prove our preceding results.

Proof of Theorem 3.4.

Note first that the tests ψ_W specified in the statement of the theorem lead to a set $\mathcal{U}$ that is closed under union, by construction. Moreover, $U = \cup_{B \in \mathcal{U}} = R_0^c$. It follows that

$$\begin{aligned} \overline{\Gamma}_{\rm inv}(C) &= \max_{B \in \mathcal{U}} \frac{\#(B \cap C)}{\#(C)} && (32) \\ &= \frac{\#(U \cap C)}{\#(C)} && (33) \\ &= \frac{\#(C \cap R_0^c)}{\#(C)} && (34) \\ &= \overline{\Gamma}_{\rm aug}(C). && (35) \end{aligned}$$

□

Proof of Theorem 4.1. Let $q_{kj} = B^{-1}_{k,j-k+1}(\alpha)$. Define

$$m^*_{0k} = m + k - \min\{k \le j \le m : P_{(j)} \ge q_{k,m-j+1}\},$$

where we take $m^*_{0k} = 0$ if the min is not satisfied. Define the maximal configuration h^*_k by

$$h^*_{ki} = \begin{cases} 0 & \text{if } P_i = P_{(j)} \text{ for some } j < k \\ 1 & \text{if } P_i = P_{(j)} \text{ for some } k \leq j < m - m^*_{0k} + k \\ 0 & \text{if } P_i = P_{(j)} \text{ for some } j \geq m - m^*_{0k} + k. \end{cases}$$

To see that the lemma holds, define $j^* = \min\{k \leq j \leq m : P_{(j)} \geq q_{k,m-j+1}\}$. Then, the vector h^*_k defined above will not be rejected by the corresponding test and has the largest pointwise FDP of any other such h vector. This is because for any other h vector for which the test is not rejected, the kth p-value with $h_i = 0$ (of rank say j in the whole vector) must be above $q_{k,\sum(1-h_i)} \geq q_{k,m-j+k}$. Hence, $P_{(j)} \geq q_{k,m-j+k}$ and $j \geq j^*$. This implies that $\Gamma(\cdot; h, P^m) \leq \Gamma(\cdot; h^*_k, P^m)$. □

Proof of Lemma 4.1. Let $k = 2$ and $m = 4$. Let q_4 and q_3 be the quantiles for the $P_{(2)}$ test on sets of size 4 and 3 respectively. Note that $q_4 < q_3$. If we have $P_{(1)} < P_{(2)} < q_4 < q_3 < P_{(3)} < P_{(4)}$. Then $U_1 = \{1, 3, 4\}$ and $U_2 = \{2, 3, 4\}$ are both in $\mathcal{U}$, but their union $\{1, 2, 3, 4\}$ is not. □

Proof of Theorem 4.2. The proof of this theorem mimics the proof of Theorem 3.4. □

Proof of Theorem 5.1. For the proof, it is sufficient to work with the envelope $\overline{\Gamma}(t)$. Define

$$H_t = \begin{cases} \dfrac{1-a}{\xi} & t \leq 1/\beta \\ \dfrac{(1-a)t}{(1-a)t + a} & t > 1/\beta. \end{cases}$$

Fix any small $\delta > 0$. From Genovese and Wasserman (2003), $\sup_{t \geq \delta} |\Gamma(t) - H_t| \to 0$, a.s. First suppose that, $(1-a)/\xi < c \leq 1 - a$. Then $H_{t_0} = c$ and $1/\beta < t_0 < 1$. By the central limit theorem and Slutsky's theorem,

$$\sqrt{m}(\Gamma(t_m) - H_{t_m}) \rightsquigarrow N(0, \sigma)$$

for some $\sigma > 0$. Hence, with $Z \sim N(0,1)$,

$$\mathbb{P}\{\Gamma(t_m) > c\} \to \mathbb{P}\{Z > z_\alpha\} = \alpha.$$

This establishes (1). If $\mathbb{P}\{T \leq T_*\} < 1-\alpha$, then a simlar calculation shows that $\mathbb{P}\{\Gamma(T) > c\} > \alpha$. This establishes (2). In the case, $c > 1-a$, $H_t < c$ for all t and so $\Gamma(1) < c$ with probability tending to 1. Both (1) and (2) follow. □

Proof of Theorem 5.2. Fix an integer k and $\epsilon > 0$. Let $\ell = 3\xi q_k/\epsilon$ where q_k is the α quantile of a Gamma$(k,1)$ distribution. Let $q(\alpha, k, m-j+1) = B^{-1}_{k,m-j+1}(\alpha)$. Then, $q(\alpha,k,m-j+1) \sim q_j/(m-j+1)$ as $m \to \infty$. Let N be the number of p-values less than or equal to $q(\alpha,k,m-\ell+1)$. Then, $N \sim$ Binomial(m,θ_m) where

$$m\theta_m = mG(q(\alpha,k,m-\ell+1)) = m\xi q(\alpha,k,m-\ell+1) \leq 2\xi q_k$$

for all large m. Thus,

$$\begin{aligned}\mathbb{P}\{J_k \geq \ell\} &= \mathbb{P}\big\{P_{(j)} \leq q(\alpha,k,m-j+1), \quad j=1,\ldots,\ell\big\}\\ &\leq \mathbb{P}\big\{P_{(j)} \leq q(\alpha,k,m-\ell+1), \quad j=1,\ldots,\ell\big\}\\ &\leq \mathbb{P}\{N \geq \ell\}\\ &\leq \frac{\mathbb{E}N}{\ell} \leq \frac{2\xi q_k}{\ell} < \epsilon.\end{aligned}$$

So, $J_k = O_P(1)$ as $m \to \infty$. From (16), T_k satisfies,

$$\widehat{G}_m(T_k) \leq \frac{J_k - k + 1}{m(1-c)}.$$

Thus, the number of observations less than or equal to T_k is no more than $(O_P(1) - k + 1)/(1-c)$, with probability tending to one. It follows that $T_k = o_P(1)$. □

Proof of Theorem 9.1. Let $\ell = |S_1|$ and let $c = c(\alpha,\ell)$. For any $F \in \mathcal{F}$, $F(t) \geq V(t)$ for all t. So,

$$\mathbb{P}_F(\phi_{S_1} = 1) = \mathbb{P}_F(\max_{i \in S_1} P_i > c) = 1 - F(c)^\ell \leq 1 - V(c)^\ell = \alpha.$$

□

Proof of Theorem 9.3. For the dependent case, note that, regardless of the dependence, $\mathbb{P}_F\{\max_{i \in S_1} P_i > c\} \leq 1 - F(c)$. Thus, replacing $c(\alpha, \ell)$ with $c(\alpha, \ell) = V^{-1}(1 - \alpha)$ yields a valid confidence envelope. □

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.

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

Efron, B., Tibshirani R., and Storey, J. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96, 1151-1160.

Finner, H. and Roters, M. (2002). Multiple hypothesis testing and expected number of type I errors. *The Annals of Statistics*, **30**, 220-238.

Genovese, C. R. and Wasserman, L. (2002). Operating Characteristics and Extensions of the False Discovery Rate Procedure, *J. Royal Statist. Soc. B*, **64**, 499–518.

Genovese, C. R. and Wasserman, L. (2004). A stochastic process approach to false discovery ccontrol. *The Annals of Statistics*, **32**, 1035-1061.

Perone Pacifico, M., Genovese, C., Verdienlli, I. and Wasserman, L. (2003). False discovery control for random fields. In press: *Journal of the American Statistical Association*,

Sarkar, S. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics*, **30**, 239-257.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479–498.

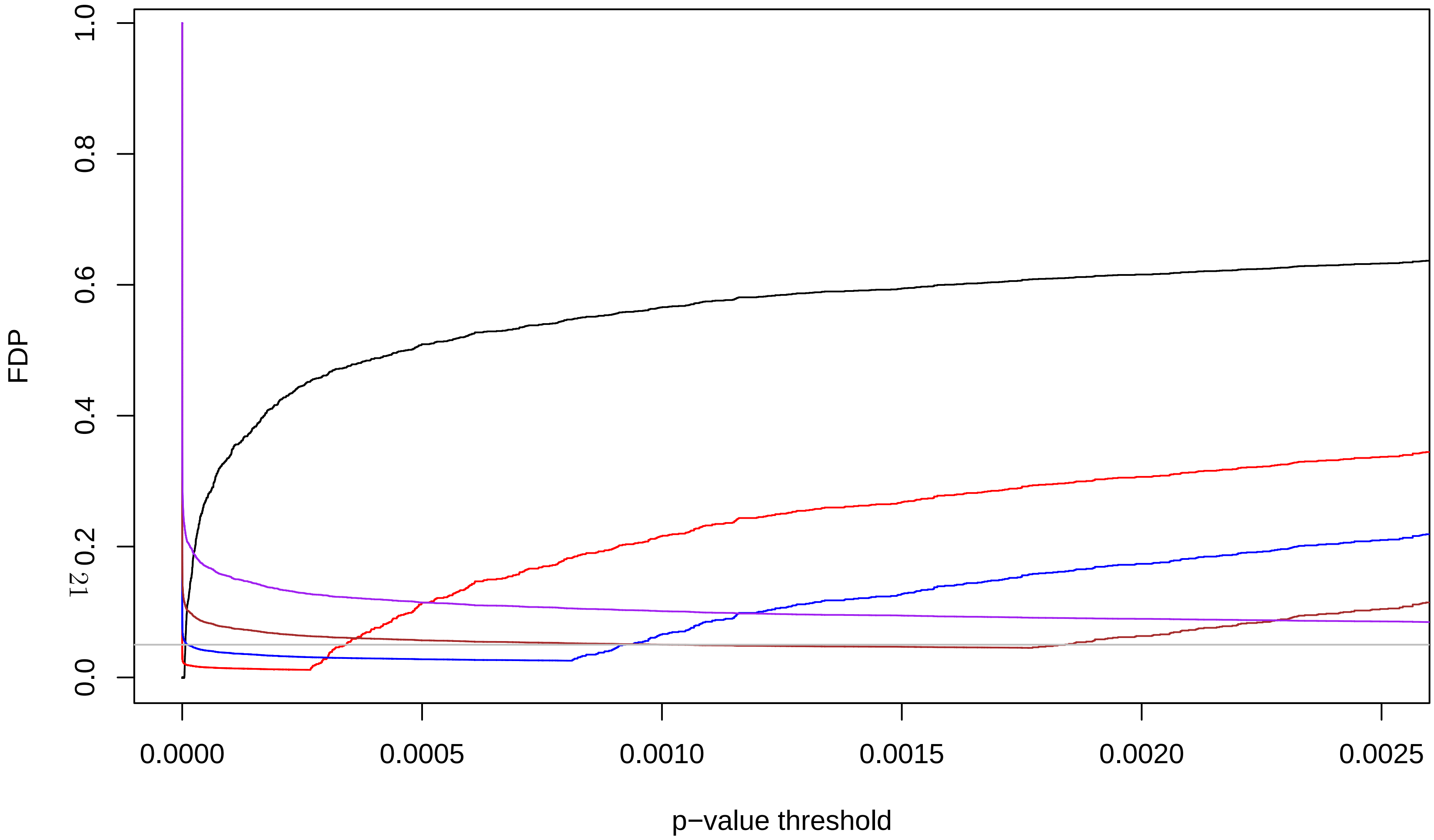
Storey, J.D. (2003). The positive False Discovery Rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, **31**, 2013–2035.

Storey J.D., Taylor J.E., and Siegmund D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, **66**, 187–205.

van der Laan, M.J., Dudoit, S., Pollard, K.S. (2004). Multiple Testing. Part III. Procedures for Control of the Generalized Family-Wise Error Rate and Proportion of False Positives, U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 141.

Figure Captions

Figure 1. $P_{(k)}$ confidence envelopes, expressed as a function of threshold, for $k = 1, 10, 25, 50, 100$. The 0.05 FDP level is marked. The $P_{(1)}$ envelope starts at 0 and strictly increases. The $P_{(k)}$ envelopes for $k > 1$ start at 1, decrease to a minimum and then increase again.


1.0
0.8
0.6
0.4
0.2
0.0
FDP
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
p−value threshold

m	a	θ	FDP Combined	Power Combined	FDP$P_{(1)}$	Power $P_{(1)}$	FDP$P_{(10)}$	Power $P_{(10)}$
100	0.01	5	0	1	0	1	0	0
	0.05	5	0.2	1	0.2	1	0	0
	0.1	5	0.2	1	0.2	1	0	0
	0.01	4	0	1	0	1	0	0
	0.05	4	0	1	0	1	0	0
	0.1	4	0.077	1	0	0.917	0	0
	0.01	3	1	0	1	0	0	0
	0.05	3	0	0.25	0	0.5	0	0
	0.1	3	0	0.5	0	0.5	0	0
	0.01	2	0	0	0	0	0	0
	0.05	2	0	0	0	0	0	0
	0.1	2	0	0.1	0	0.2	0	0
1000	0.01	5	0.091	1	0.167	1	0	0
	0.05	5	0.183	1	0.14	1	0.183	1
	0.1	5	0.162	1	0.101	1	0.173	1
	0.01	4	0.286	0.5	0.286	0.5	0	0
	0.05	4	0.151	0.957	0	0.596	0.182	0.957
	0.1	4	0.12	0.957	0	0.707	0.137	0.957
	0.01	3	0	0	0	0	0	0
	0.05	3	0.114	0.62	0	0.34	0	0
	0.1	3	0.104	0.674	0	0.281	0.113	0.708
	0.01	2	0	0	0	0	0	0
	0.05	2	0	0.016	0	0.032	0	0
	0.1	2	0	0.07	0	0.05	0	0
10000	0.01	5	0.102	0.98	0	0.889	0.118	0.98
	0.05	5	0.179	0.994	0.004	0.917	0.172	0.994
	0.1	5	0.178	0.998	0.001	0.905	0.162	0.997
	0.01	4	0.08	0.741	0.022	0.407	0.109	0.759
	0.05	4	0.125	0.95	0	0.424	0.045	0.887
	0.1	4	0.164	0.974	0.002	0.436	0.044	0.915
	0.01	3	0	0.265	0	0.098	0	0
	0.05	3	0.127	0.623	0	0.106	0.05	0.463
	0.1	3	0.137	0.79	0	0.087	0.018	0.472
	0.01	2	0	0	0	0.01	0	0

Table 1: Simulation Results