Chapter 1

Adaptive Multiple Testing Procedures under Positive Dependence

Wenge Guo, Sanat K. Sarkar and Shyamal D. Peddada*

Department of Mathematical Sciences, New Jersey Institute of Technology
Newark, New Jersey, 07102, U.S.A.
wenge.guo@njit.edu

Department of Statistics, Temple University
Philadelphia, PA 19122, U.S.A.
sanat@temple.edu

Biostatistics Branch, National Institute of Environmental Health Sciences
Research Triangle Park, NC 27709, U.S.A.
peddada@niehs.nih.gov

In multiple testing, the unknown proportion of true null hypotheses among all null hypotheses that are tested often plays an important role. In adaptive procedures this proportion is estimated and then used to derive more powerful multiple testing procedures. Hochberg and Benjamini (1990) first presented adaptive Holm and Hochberg procedures for controlling the familywise error rate (FWER). However, until now, no mathematical proof has been provided to demonstrate that these procedures control the FWER in finite samples. In this paper, we present new adaptive Holm and Hochberg procedures and prove they can control the FWER in finite samples under some common types of positive dependence. Through a small simulation study, we illustrate that these adaptive procedures are more powerful than the corresponding non-adaptive procedures.

*The research of Wenge Guo is supported by NSF Grants DMS-1006021, the research of Sanat Sarkar is supported by NSF Grants DMS-0603868 and DMS-1006344, and the research of Shyamal Peddada is supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101744-04). The authors thank Gregg E. Dinse, Mengyuan Xu and the referee for carefully reading of this manuscript and for their useful comments that improved the presentation.

1.1. Introduction

In this article, we consider the problem of simultaneously testing a finite number of null hypotheses H_i, $i = 1, \ldots, n$, based on their respective p-values P_i, $i = 1, \ldots, n$. A main concern in multiple testing is the multiplicity problem, namely, that the probability of committing at least one Type I error sharply increases with the number of hypotheses tested at a pre-specified level. There are two general approaches for dealing with this problem. The first one is to control the familywise error rate (FWER), which is the probability of one or more false rejections, and the second one is to control the false discovery rate (FDR), which is the expected proportion of Type I errors among the rejected hypotheses, proposed by Benjamini and Hochberg (1995). The first approach works well for traditional small-scale multiple testing , while the second one is more suitable for modern large-scale multiple testing problems.

Given the ordered p-values $P_{1:n} \leq \cdots \leq P_{n:n}$ with the associated null hypotheses $H_{1:n}, \cdots, H_{n:n}$, and a non-decreasing sequence of critical values $\alpha_1 \leq \cdots \leq \alpha_n$, there are two main avenues open for developing multiple testing procedures based on the marginal p-values – stepdown and stepup.

- A stepdown procedure based on these critical values operates as follows. If $P_{1:n} > \alpha_1$, do not reject any hypothesis. Otherwise, reject hypotheses $H_{1:n}, \cdots, H_{r:n}$, where $r \geq 1$ is the largest index satisfying $P_{1:n} \leq \alpha_1, \cdots, P_{r:n} \leq \alpha_r$. If, however, $P_{r:n} > \alpha_r$ for all $r \geq 1$, then do not reject any hypothesis. Thus, a stepdown procedure starts with the most significant hypothesis and continues rejecting hypotheses as long as their corresponding p-values are less than or equal to the corresponding critical values.
- A stepup procedure , on the other hand, operates as follows. If $P_{n:n} \leq \alpha_n$, then reject all null hypotheses; otherwise, reject hypotheses $H_{1:n}, \cdots, H_{r:n}$, where $r \geq 1$ is the smallest index satisfying $P_{n:n} > \alpha_n, \ldots, P_{r+1:n} > \alpha_{r+1}$. If, however, $P_{r:n} > \alpha_r$ for all $r \geq 1$, then do not reject any hypothesis. Thus, a stepup procedure begins with the least significant hypothesis and continues accepting hypotheses as long as their corresponding p-values are greater than the corresponding critical values until reaching the most significant hypothesis $H_{1:n}$.

If $\alpha_1 = \cdots = \alpha_n$, the stepup or stepdown procedure reduces to what is usually referred to as a single-step procedure.

For controlling the FWER, a number of widely used procedures are available, among which the Bonferroni, Holm (1979) and Hochberg (1988) procedures are relatively popular. The Bonferroni procedure is a single-step procedure with the critical values $\alpha_i = \alpha/n, i = 1, \ldots, n$. The Holm procedure is a stepdown procedure with the critical values $\alpha_i = \alpha/(n - i + 1), i = 1, \ldots, n$, and the Hochberg procedure is a stepup procedure based on the same set of critical values as Holm's. With the null p-values having the $U(0, 1)$, or stochastically larger than the $U(0, 1)$, distribution, the Bonferroni and Holm procedures both control the FWER at α without any further assumption on the dependence structure of the p-values. The Hochberg procedure controls the FWER at α when the null p-values are independent or positively dependent in the the following sense:

$$E\{\phi(P_1, \ldots, P_n) \mid P_i = u\} \uparrow u \in (0, 1), \tag{1.1}$$

for each P_i and any increasing (coordinatewise) function ϕ. (Hochberg, 1988; Sarkar, 1998; Sarkar & Chang, 1997). The condition (1) is the positive dependence through stochastic ordering (PDS) condition defined by Block, Savits and Shaked (1985), although it is often referred to as the positive regression dependence on subset (of null p-values), the PRDS condition, considered in Benjamini & Yekutieli (2001) and Sarkar (2002) in the context of FDR . Also, it has been noted recently that this positive dependence condition can be replaced by the following weaker condition:

$$E\{\phi(P_1, \ldots, P_n) \mid P_i \leq u\} \uparrow u \in (0, 1). \tag{1.2}$$

The condition (1) or (2) is satisfied by a number of multivariate distributions arising in many multiple testing situations, for example, those of multivariate normal test statistics with positive correlations, absolute values of studentized independent normals, and multivariate t and F (Benjamini & Yekutieli, 2001; Sarkar, 2002).

Since these procedures are often conservative by a factor which is the unknown proportion of true null hypotheses, the conservativeness in these procedures could be reduced, and hence the power could potentially be increased, if an estimate of this proportion can be suitably incorporated into these procedures. With that idea in mind, Hochberg & Benjamini (1990) proposed adaptive Bonferroni , Holm and Hochberg procedures for controlling the FWER. However, it has not been proved yet that these adaptive FWER procedures actually can provide an ultimate control over the FWER. Recently, Guo (2009) introduced new adaptive Bonferroni and Holm procedures by simplifying those in Hochberg & Benjamini (1990). He

proved that, under a conditional independence model, while his adaptive Bonferroni procedure controls the FWER for finite samples, the adaptive Holm procedure approximately controls the FWER for large samples.

For controlling the FDR, the well-known procedure is that of Benjamini and Hochberg (1995). The same phenomenon in terms of conservativeness happens also with this procedure and a number of adaptive versions of it that control the FDR have been introduced in the literature; see Benjamini & Hochberg (2000), Storey et al. (2004), Benjamini et al. (2006), Ferreira & Zwinderman (2006), Sarkar (2006, 2009), Benjamini & Heller (2007), Farcomeni (2007), Blanchard & Roquain (2008), Wu (2008), Gavrilov et al. (2009), and Sarkar & Guo (2009). It is important to note that in the case of finite samples, the FDR control of these adaptive procedures has been proved only when the underlying test statistics are independent. Using a simulation study, Benjamini et al. (2006) demonstrated that some adaptive FDR procedures, such as Storey's, which control the FDR under independence, may fail to do so under dependence. Thus, developing an adaptive procedure controlling the FWER or FDR even under dependence in finite samples appears to be an important undertaking.

In this paper, we concentrate mainly on developing adaptive FWER procedures. We take a general approach to constructing such a procedure that controls the FWER under independence or positive dependence. This involves using a concept of adaptive global testing and the closure principle of Marcus et al. (1976). The closure principle is a useful tool to derive FWER controlling multiple testing procedures based on valid tests available for different possible intersections or global null hypotheses. In adaptive global testing, information about the number of true null hypotheses is extracted from the available p-values and incorporated into a procedure while testing an intersection or global null hypothesis and maintaining a control over the (global) type I error rate. We derive two such adaptive global tests, with one involving an estimate of the number of true null hypotheses considered in Hommel's (1988) FWER controlling procedure and the other based on an estimate of this number that can be obtained by applying the Benjamini and Hochberg's (1995) FDR controlling procedure. Both of these tests provide valid controls of the (global) type I error rate under independence or positive dependence, in the sense of (1) or (2), of the p-values. Based on these global tests and applying the closure principle, we derive alternative adaptive Holm and adaptive Hochberg procedures. We offer theoretical proofs of the FWER controls of these procedures in finite samples under independence or positive dependence in the sense of (1) or

(2) of the p-values. We provide numerical evidence through a small-scale simulation study that the present adaptive Holm and Hochberg procedures can be more powerful, as expected, than the corresponding non-adaptive procedures.

The paper is organized as follows. In Section 2, we introduce what we mean by an adaptive global test and present two such tests. In Section 3, we present the developments of our proposed new adaptive Holm and Hochberg procedures and prove that they control the FWER under independence or positive dependence in the sense of (1) or (2) of the p-values. A real-life application of our procedures and the results of a simulation study investigating the performances of our procedures relative to others are also presented in this section. Some concluding remarks are made in Section 4.

1.2. Adaptive Global Tests

In this section, we will present our idea of an adaptive global test . Given any family of null hypothesis $H_1, \ldots, H_m$, and the corresponding p-values P_i, $i = 1, \ldots, m$, consider testing the global null hypothesis $H_0 = \bigcap_{i=1}^{m} H_i$. We will focus on global tests where the rejection regions are of the form $\bigcup_{i=1}^{m} \{P_{i:m} \leq c_i\}$; that is, where each ordered p-value $P_{i:m}$ is compared with a cut-off point c_i, with $0 \leq c_1 \leq \ldots \leq c_m \leq 1$, and H_0 is rejected if $P_{i:m} \leq c_i$ holds for at least one i. Such a test has been referred to as a cut-off test (Bernhard et al., 2004). It allows making decisions on the individual null hypotheses once the global null hypothesis is rejected, which is important since we need to develop in the next section multiple testing procedures based on it through the closure principle.

There are a number of such cut-off global tests available in the literature, such as the Bonferroni test , where $c_1 = \ldots = c_m = \alpha/m$, and the Simes test, where $c_i = i\alpha/m$, for $i = 1, \ldots, m$, (Simes, 1986). However, the idea of extracting information about the number, say m_0, of the true null hypotheses in the family of interest and incorporating that into the construction of a global cut-off test has net yet been seen in the literature. Why would such an adaptive global test make sense? Consider, for instance, the statistic $W_m(\lambda) = \sum_{i=1}^{m} I(P_i > \lambda)$ (with I being the indicator function), which is the number of insignificant p-values observed when the fixed rejection threshold $\lambda \in (0, 1)$ is chosen for each p-value. A high value of $W_m(\lambda)$ would indicate that m_0 is likely to be large, and hence would provide an evidence towards accepting the global null hypothesis. Similarly, a small value of $W_m(\lambda)$ would provide an evidence towards re-

jecting the null hypothesis. It is important to note that, although a test just based on $W_m(\lambda)$, for a fixed λ and controlling the type I error rate at a pre-specified level could be formulated for testing H_0, either exactly using the Binomial distribution when the p-values are independent, since in this case $W_m(\lambda) \sim \text{Bin}(m, 1-\lambda)$ under H_0, or approximately using, for example, a permutation test when the p-values have an unknown or more complicated dependence structure, this would not be helpful in terms of providing a cut-off test. Nevertheless, the value of $W_m(\lambda)$ can be factored into each P_i in a way that shrinks P_i towards a smaller value, making it more likely to be significant, if $W_m(\lambda)$ is small, and expands P_i to a larger value if $W_m(\lambda)$ is large. Of course, instead of $W_m(\lambda)$, we could use any other statistic, or a consistent estimate of m_0, a large value of which would indicate acceptance of H_0. This is how we will develop two global cut-off tests in the following.

First, we develop our adaptive Simes global test , borrowing the idea of estimating m_0 from Benjamini et al. (2006). Let $\mathbf{P}_m = (P_1, \ldots, P_m)$, and $W_1(\mathbf{P}_m)$ be the total number of accepted null hypotheses when the FDR controlling procedure of Benjamini and Hochberg (1995), the BH procedure, is applied to $\mathbf{P}_m$. Recall that the BH procedure is a stepup procedure based on the critical values of the original Simes global test . With

$$\widehat{m}_0^{(1)}(\mathbf{P}_m) = \max\{W_1(\mathbf{P}_m), 1\}, \tag{2.3}$$

we define the following:

Adaptive Simes Test. *Reject H_0 if $P_{i:m} \le c_i$ for at least one $i = 1, \ldots, m$, where $c_i = i\alpha/\widehat{m}_0^{(1)}(\mathbf{P}_m)$.*

The fact that the adaptive Simes test controls the type I error rate at α under independence or positive dependence in the sense of (1) or (2) can be proved as follows. Let R_1 and R_2 be the total numbers of the null hypotheses rejected by the BH procedure and the stepup procedure with the same critical values as in the adaptive Simes test. Then, the type I error rate of the adaptive Simes test is give by

$$\text{pr}\{R_2 > 0\} = \text{pr}\{R_2 > 0, R_1 = 0\} + \text{pr}\{R_2 > 0, R_1 > 0\}, \tag{2.4}$$

with the probabilities being evaluated under H_0. Since $\widehat{m}_0^{(1)}(\mathbf{P}_m) = \max\{m - R_1, 1\}$, and $R_2 = 0$ with probability one if $R_1 = 0$,

$$\text{pr}\{R_2 > 0\} = \text{pr}\{R_2 > 0, R_1 > 0\} \le \text{pr}\{R_1 > 0\},$$

which is less than or equal to α under independence or positive dependence in the sense of (1) or (2) of the p-values due to the well-known Simes' inequality (Simes, 1986; Sarkar, 1998; Sarkar & Chang, 1997).

We will now obtain an adaptive Bonferroni global test. We will do that through reinterpreting the Hommel procedure (Hommel, 1988) as an adaptive version of the Bonferroni procedure. The Hommel procedure is defined as follows. Let

$$W_2(\mathbf{P}_m) = \{j \in \{1, \ldots, m\} : P_{m-j+k:m} > k\alpha/j,\ k = 1, \ldots, j\}$$

and

$$\widehat{m}_0^{(2)}(\mathbf{P}_m) = \max\{W_2(\mathbf{P}_m), 1\}. \tag{2.5}$$

If $W_2(\mathbf{P}_m)$ is nonempty, reject H_i whenever $P_i \leq \alpha/\widehat{m}_0^{(2)}(\mathbf{P}_m)$. If, however, $W_2(\mathbf{P}_m)$ is empty, reject all H_i, $i = 1, \ldots, m$. Notice that $\widehat{m}_0^{(2)}(\mathbf{P}_m)$ represents the maximum size of the subfamily of null hypotheses whose members are all declared to be true when applying the Simes test. In other words, $\widehat{m}_0^{(2)}(\mathbf{P}_m)$ provides an estimate of m_0, in terms of which the Hommel procedure can be interpreted as the following:

Adaptive Bonferroni Test. *Reject H_0 if $P_{i:m} \leq c_i$ for at least one $i = 1, \ldots, m$, where $c_i = \alpha/\widehat{m}_0^{(2)}(\mathbf{P}_m)$.*

We can summarize the above discussions in the following proposition.

Proposition 1.1. *Given any family of null hypotheses H_i, $i = 1, \ldots, m$, consider testing the global null hypothesis $H_0 = \bigcap_{i=1}^m H_i$ using a cut-off test of the form $\bigcup_{i=1}^m \{P_{i:m} \leq c_i\}$. When the p-values are independent or positively dependent in the sense of (1) or (2), the adaptive Simes and Bonferroni tests are valid level α tests.*

Remark 1.1. In the above adaptive tests, we use $\max\{1, i\}$ as the estimate of m_0 once the hypotheses $H_{m-i+k:m}$, $k = 1, \ldots, i$, are accepted, where this i, for the adaptive Simes test, is the index such that $P_{m-i+k:m} > (m - i + k)\alpha/m$ for all $k = 1, \ldots, i$, whereas, for the adaptive Bonferroni test, it is the largest index from 1 to m such that $P_{m-i+k:m} > k\alpha/i$ for all $k = 1, \ldots, i$. Since $(m - i + k)\alpha/m \geq k\alpha/i$ for $k = 1, \ldots, i$, the estimate of m_0 is more liberal in the adaptive Simes test than in the adaptive Hommel test, i.e., $\hat{m}_0^{(1)} \leq \hat{m}_0^{(2)}$, implying that the adaptive Simes test is more powerful.

Remark 1.2. In the alternative adaptive Bonferroni procedure considered in Guo (2009), $c_i = \alpha/\hat{m}_0$, $i = 1, \ldots, m$, where $\widehat{m}_0 = [W_m(\lambda) + 1]/(1 - \lambda)$. It also provides a valid level α global test for testing H_0, but under a model that assumes independence of the p-values conditional on any (random) configurations of true and false null hypotheses.

1.3. Adaptive Multiple Testing Procedures

We now consider our main problem, which is, to simultaneously test the null hypotheses H_i, $i = 1, \ldots, n$, and develop newer adaptive versions of Holm's stepdown and Hochberg's stepup procedures that utilize information about the number of true null hypotheses suitably extracted from the data and ultimately maintain a control over the FWER at α. We first present these procedures. Then, we provide a real life application, and the results of a simulation study investigating the performances of our proposed adaptive procedures in relation to those of the corresponding conventional, non-adaptive procedures.

1.3.1. *The Procedures*

We will develop our procedures using the following closure principle of Marcus et al. (1976) that is often used to construct FWER controlling procedures.

Closure Principle. *Suppose that for each $I \subseteq \{1, \ldots, n\}$ there is a valid level α global test for testing the intersection null hypothesis $\bigcap_{i \in I} H_i$. An individual null hypothesis H_i is rejected if for each $I \subseteq \{1, \ldots, n\}$ with $I \ni i$, $\bigcap_{j \in I} H_j$ is rejected by the corresponding global test.*

A multiple testing procedure satisfying the closure principle is termed a closed testing procedure. It controls the FWER at α. Many of the multiple testing procedures in the literature controlling the FWER are either closed testing procedures or can be presented as some versions of such a procedure. The level α adaptive global tests presented in the above section will be the key towards developing our proposed adaptive FWER controlling procedures based on the closure principle. Before we do that, we first need to introduce a few more additional notations.

Consider all possible sub-families of the null hypotheses, $\{H_i,\ i \in I_m\}$, $I_m \subseteq \{1, \ldots, n\}$, $m = 1, \ldots, n$, where I_m is of cardinality m. Define $\widehat{n}_0^{(1)}(\mathbf{P}_m)$ and $\widehat{n}_0^{(2)}(\mathbf{P}_m)$, the two estimates of the number of true null hypotheses in $\{H_i, i \in I_m\}$ based on the corresponding set of p-values $\mathbf{P}_m = \{P_i, i \in I_m\}$, as in (3) and (5), respectively. Since these estimates are symmetric and componentwise increasing in $\mathbf{P}_m$, and every ordered component of any m-dimensional subset of the p-values is smaller than the corresponding component of $\widetilde{\mathbf{P}}_m = (P_{n-m+1:n}, \ldots, P_{n:n})$, we have the following: $\widehat{n}_0^{(j)}(\widetilde{\mathbf{P}}_m) \geq \widehat{n}_0^{(j)}(\mathbf{P}_m)$, for any $\mathbf{P}_m$ and $j = 1, 2$. For convenience, we will denote $\widehat{n}_0^{(j)}(\widetilde{\mathbf{P}}_m)$ simply as $\widehat{n}_0^{(j)}(m)$, $j = 1, 2$.

It is important to note exactly how $\widehat{n}_0^{(j)}(m)$ is defined for $j = 1, 2$. Consider using the p-values $P_{n-m+1:n}, \ldots, P_{n:n}$ to test corresponding null hypotheses. Then, from (3), $\widehat{n}_0^{(1)}(m) = \max\{W_1(\widetilde{\mathbf{P}}_m), 1\}$, where $W_1(\widetilde{\mathbf{P}}_m)$ is the number of accepted null hypotheses in the stepup test involving these p-values and the critical values $j\alpha/m$, $j = 1, \ldots, m$. Similarly, from (5), $\widehat{n}_0^{(2)}(m) = \max\{W_2(\widetilde{\mathbf{P}}_m), 1\}$, where

$$W_2(\widetilde{\mathbf{P}}_m) = \{j \in \{1, \ldots, m\} : P_{n-j+k:n} > k\alpha/j,\ k = 1, \ldots, j\}.$$

It is easy to see that while $\widehat{n}_0^{(2)}(m)$ is increasing in m, $\widehat{n}_0^{(1)}(m)$ may not be so. If $\widehat{n}_0^{(1)}(m)$ is not increasing in m, we will make a minor modification of it as follows: Let $\widehat{n}_0^{(1)'}(1) = \widehat{n}_0^{(1)}(1)$ and $\widehat{n}_0^{(1)'}(m) = \max\left(\widehat{n}_0^{(1)'}(m-1), \widehat{n}_0^{(1)}(m)\right)$, for $2 \le m \le n$. Obviously, such modified $\widehat{n}_0^{(1)'}(m)$ is always increasing in m, and for each $m = 1, \ldots, n$, $\widehat{n}_0^{(1)'}(m) \ge \widehat{n}_0^{(1)}(m)$, with the equality holding when $\widehat{n}_0^{(1)}(m)$ is increasing in m.

Now, we present our adaptive Holm procedure in the following theorem.

Theorem 1.1. *Consider the the stepdown procedure with the critical values* $\alpha/\hat{n}_0^{(2)}(n-j+1)$, $j = 1, \ldots, n$. *It controls the FWER at* α *when the p-values are independent or positively dependent in the sense of (1) or (2).*

Proof. Suppose that $P_{j:n}$ is the smallest among the p-values that correspond to the n_0 true null hypotheses. If $P_{j:n} \le \alpha/\hat{n}_0^{(2)}(n-j+1)$, then for any m-dimensional subset of the null hypotheses containing the true null hypothesis corresponding to $P_{j:n}$, the adaptive Bonferroni test with the critical constants $c_j = \alpha/\hat{n}_0^{(2)}(\mathbf{P}_m)$ rejects its intersection H_0^m, where $\mathbf{P}_m$ is the corresponding p-value vector of the m individual null hypotheses. Since under H_0^m, $m \le n_0 \le n-j+1$, then we have $P_{j:n} \le \alpha/\hat{n}_0^{(2)}(n-j+1) \le \alpha/\hat{n}_0^{(2)}(m) \le \alpha/\hat{n}_0^{(2)}(\mathbf{P}_m)$. Thus, if $P_{j:n} \le \alpha/\hat{n}_0^{(2)}(n-j+1)$, $H_{j:n}$ is rejected by the closed testing procedure based on the above Bonferroni test. Therefore, $\mathrm{pr}\{P_{j:n} \le \alpha/\hat{n}_0^{(2)}(n-j+1)\}$ is less than or equal to the FWER of the closed testing procedure. By the closure principle and Proposition 1, $\mathrm{pr}\{P_{j:n} \le \alpha/\hat{n}_0^{(2)}(n-j+1)\} \le \alpha$. Therefore, the FWER of the adaptive Holm procedure is less than or equal to α. ■

Remark 1.3. In the alternative adaptive Holm's procedure of Guo (2009), $c_i = \alpha/\min\{n-i+1, \hat{n}_0\}$, $i = 1, \ldots, n$, where $\widehat{n}_0 = [W_n(\lambda)+1]/(1-\lambda)$. It asymptotically (as $n \to \infty$) controls the FWER at α under a conditional independence model (Wu, 2008). The adaptive Holm procedure in Theorem

1, on the other hand, not only controls the FWER in finite samples but also under a more general type of dependence situation.

Next, we present our adaptive Hochberg procedure through the adaptive Simes test defined in the preceding section.

Theorem 1.2. *Consider the stepup procedure with the critical values* $\alpha/\hat{n}_0^{(1)'}(n-j+1), j=1,\ldots,n$. *It controls the FWER at* α *when the p-values are independent or positively dependent in the sense of (1) or (2).*

PROOF. Let $i_0 = \max\{i : P_{i:n} \le \hat{n}_0^{(1)'}(n-i+1)\}$. First, for any subset of m individual hypotheses such that the corresponding smallest p-value is $P_{i_0:n}$, the adaptive Simes test with the critical constants $c_j = j\alpha/\hat{n}_0^{(1)}(\mathbf{P}_m)$, $j=1,\ldots,m$ rejects its intersection hypothesis, since $m \le n-i_0+1$ and thus $P_{i_0:n} \le \alpha/\hat{n}_0^{(1)'}(n-i_0+1) \le \alpha/\hat{n}_0^{(1)'}(m) \le \alpha/\hat{n}_0^{(1)}(m) \le \alpha/\hat{n}_0^{(1)}(\mathbf{P}_m)$, where $\mathbf{P}_m$ is the corresponding p-value vector of the m hypotheses. Second, consider a different subset of m individual hypotheses with exactly k hypotheses whose p-value is less than $P_{i_0:n}$. It is easy to see that $\hat{n}_0^{(1)'}(j+1) \le \hat{n}_0^{(1)'}(j)+1$ for any $1 \le j \le n-1$, thus $\hat{n}_0^{(1)'}(n-i_0+1+k) \le \hat{n}_0^{(1)'}(n-i_0+1)+k$. Also, $m \le n-i_0+1+k$. Therefore,

$$\begin{aligned} P_{i_0:n} &\le \frac{\alpha}{\hat{n}_0^{(1)'}(n-i_0+1)} \le \frac{(k+1)\alpha}{\hat{n}_0^{(1)'}(n-i_0+1)+k} \\ &\le \frac{(k+1)\alpha}{\hat{n}_0^{(1)'}(n-i_0+1+k)} \le \frac{(k+1)\alpha}{\hat{n}_0^{(1)'}(m)} \\ &\le \frac{(k+1)\alpha}{\hat{n}_0^{(1)}(m)} \le \frac{(k+1)\alpha}{\hat{n}_0^{(1)}(\mathbf{P}_m)}. \end{aligned} \tag{3.6}$$

In (6), the second inequality follows from the fact that $(k+1)\alpha/\hat{n}_0^{(1)'}(n-i_0+1)+k$ is increasing in k. Thus, in such situation, the adaptive Simes test also rejects its intersection hypothesis. Summarizing these two cases, the closed testing procedure based on the adaptive Simes tests will reject $H_{i_0:n}$.

For other null hypothesis $H_{i:n}$ with $i < i_0$, we only need to prove that for each subset of m individual hypotheses without containing $H_{i_0:n}$ for which $P_{i:n}$ is the $(k+1)$-smallest p-value less than $P_{i_0:n}$, its intersection hypothesis is rejected by the adaptive Simes test. Actually, by using the same arguments as (6), we can prove that $P_{i:n} \le (k+1)\alpha/\hat{n}_0^{(1)}(\mathbf{P}_m)$. Thus, $H_{i:n}$ is also rejected by the closed testing procedure. By using the closure principle and proposition 1, the adaptive Hochberg procedure controls the

FWER at level α when the p-values are independent or positively dependent in the sense of (1) or (2). ■

Remark 1.4. It is easy to see that for each $1 \leq i \leq n$, $\hat{n}_0^{(1)}(n-i+1) \leq n-i+1$, thus $\hat{n}_0^{(1)'}(n-i+1) \leq n-i+1$. Therefore, the adaptive Hochberg procedure is more powerful than the corresponding non-adaptive one.

1.3.2. *An Application*

We revisit a dose-finding diabetes trial study analyzed in Dmitrienko et al. (2007). The trial compares three doses of an experimental drug versus placebo. The efficacy profile of the drug was studied using three endpoints: Haemoglobin A1c (primary), Fasting serum glucose (secondary), and HDL cholesterol (secondary). These endpoints were examined at each of the three doses, and the raw p-values are 0.005, 0.011, 0.018, 0.009, 0.026, 0.013, 0.010, 0.006, and 0.051. We pre-specify $\alpha = 0.05$. By using the conventional, non-adaptive Holm and Hochberg procedures, we see that two null hypotheses are rejected at level 0.05 for both these tests. In contrast, our proposed adaptive Holm and Hochberg procedures both reject seven null hypotheses at the same level.

1.3.3. *A Simulation Study*

We performed a small scale simulation study investigating the performances of our proposed adaptive Holm and Hochberg procedures in comparison with those of the corresponding conventional, non-adaptive Holm and Hochberg procedures. We made these comparisons in terms of the FWER control at the desired level and power, with the power being defined as the expected proportion of the false null hypotheses that are correctly rejected.

We generated $n = 50$ dependent normal random variables $N(\mu_i, 1)$, $i = 1, \ldots, n$, with a common correlation $\rho = 0.2$, and with n_0 of the 50 μ_i's being equal to 0 and the remaining equal to 3, and applied the four different procedures to test $H_i : \mu_i = 0$ against $K_i : \mu_i \neq 0$ simultaneously for $i = 1, \ldots, 50$ at level $\alpha = 0.05$. We repeated these steps $2,000$ times before calculating the proportion (estimated FWER) of times at least one true null hypothesis is falsely rejected and the average proportion (estimated power) of false null hypotheses that are rejected. Figures 1 and 2 present the estimated FWERs and powers, respectively, of the four procedures, each plotted against different values of n_0. As seen from Figure 1, our suggested adaptive Holm and Hochberg procedures provide better control of the FWER than those of

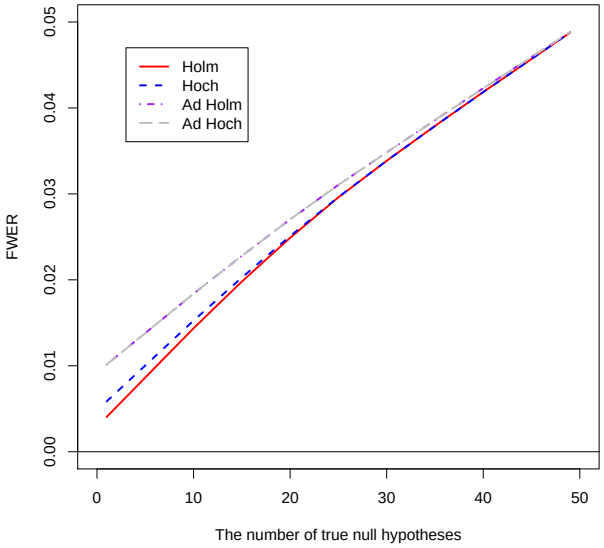

Fig. 1.1. Comparison of familywise error rates of four procedures: Holm (solid), Hochberg (small dashes), adaptive Holm (dot-dash), and adaptive Hochberg (dashes), with parameters $n = 50$, $\alpha = 0.05$.

the conventional, non-adaptive Holm and Hochberg procedures, although with increasing number of true null hypotheses all procedures become less and less conservative. Figure 2 presents the comparisons in terms of power. As seen from this figure, our suggested adaptive Holm and Hochberg procedures have better power performances than the corresponding non-adaptive Holm and Hochberg procedures. Again, with increasing number of true null hypotheses, the difference in power gets smaller and closer to zero.

1.4. Concluding Remarks

A knowledge of the proportion of true null hypotheses among all the null hypotheses tested can be useful for developing improved versions of conventional FDR or FWER controlling procedures. A number of adaptive versions of an FDR or FWER controlling procedure exist in the literature, each attempts to improve the FDR or FWER procedure by extracting information about the number of true null hypotheses from the available

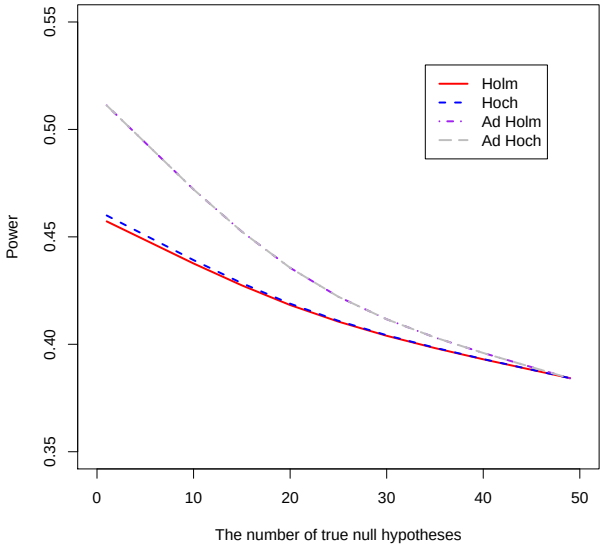

Fig. 1.2. Comparison of average power of four procedures: Holm (solid), Hochberg (small dashes), adaptive Holm (dot-dash), and adaptive Hochberg (dashes), with parameters $n = 50$, $\alpha = 0.05$.

data and incorporating that into the procedure. However, in finite sample settings, the ultimate control of the FDR or FWER for these adaptive procedures has only been proved under the assumption of independence or conditional independence of the p-values. In this article, we make an attempt for the first time, as far as we know, to develop adaptive FWER procedures that provide ultimate control over the FWER not only under independence but also under positive dependence of the p-values.

It is important to point out that there are some essential differences between adaptive and non-adaptive procedures. For example, for a non-adaptive single-step FWER controlling procedure, weak control implies strong control, but that conclusion does not hold for an adaptive single-step procedure. We explain that phenomenon through the following example.

Example 1.1. Consider an adaptive Bonferroni procedure for which $\hat{n}_0^{(1)}(n)$ is used as the estimate of the number of true null hypotheses. For convenience, we will denote $\hat{n}_0^{(1)}(n)$ simply as $\hat{n}_0^{(1)}$. By proposition 1, the

single-step procedure can weakly control the FWER. However, we can show that it cannot strongly control the FWER. Let $n = 6$ and $n_0 = 2$. Suppose four false null p-values are zero and two true null p-values q_1 and q_2 are independent identically distributed $U(0,1)$. Let $q_{(1)} \leq q_{(2)}$ denote the ordered values of q_1 and q_2. Let R be the total number of rejections. Then,

$$\text{FWER} = \text{pr}\{q_{(1)} \leq \alpha/\hat{n}_0^{(1)}, 4 \leq R \leq 6\}.$$

Note that

$$\begin{aligned}
&\text{pr}\{q_{(1)} \leq \alpha/\hat{n}_0^{(1)}, R = 4\} = \text{pr}\{q_{(1)} \leq \alpha/2, q_{(1)} > 5\alpha/6, q_{(2)} > \alpha\} = 0,\\
&\text{pr}\{q_{(1)} \leq \alpha/\hat{n}_0^{(1)}, R = 5\} = \text{pr}\{q_{(1)} \leq \alpha, q_{(1)} \leq 5\alpha/6, q_{(2)} > \alpha\}\\
&\quad = \text{pr}\{q_{(1)} \leq 5\alpha/6, q_{(2)} > \alpha\},
\end{aligned}$$

and

$$\begin{aligned}
&\text{pr}\{q_{(1)} \leq \alpha/\hat{n}_0^{(1)}, R = 6\} = \text{pr}\{q_{(1)} \leq \alpha, q_{(2)} \leq \alpha\}\\
&\quad \geq \text{pr}\{q_{(1)} \leq 5\alpha/6, q_{(2)} \leq \alpha\}.
\end{aligned}$$

Thus

$$\text{FWER} = \sum_{r=4}^{6} \text{pr}\{q_{(1)} \leq \alpha/\hat{n}_0^{(1)}, R = r\} \geq \text{pr}\{q_{(1)} \leq 5\alpha/6\} > \alpha.$$

References

[1] Benjamini, Y. and Heller, R. (2007). False discovery rate for spatial signals. *J. Am. Statist. Assoc.* **102**, 1272–1281.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300.

[3] Benjamini, Y. and Hochberg, Y. (2000). The adaptive control of the false discovery rate in multiple hypothesis testing with independent statiatics. *J. Educ. Behav. Statist.* **25**, 60–83.

[4] Benjamini, Y., Krieger, K. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507.

[5] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–1188.

[6] Bernhard, G., Klein, M. and Hommel, G. (2004). Global and multiple test procedures using ordered p-values – A review. *Statist. Pap.* **45**, 1-14.

[7] Blanchard, G. and Roquain, E. (2009). Adaptive FDR control under independence and dependence. *Journal of Machine Learning Research* **10**, 2837-2871.

[8] Block, H. W., Savits, T. H. and Shaked, M. (1985). A concept of negative dependence using stochastic ordering. *Statist. Probab. Lett.* **3**, 81-86.

[9] Dmitrienko, A., Wiens, B., Tamhane, A. and Wang, X. (2007). Global and Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statist. Med.* **26**, 2465–2478.
[10] Farcomeni, A. (2007). Some results on the control of the false diiscovery rate under dependence. *Scandinavian Journal of Statistics* **34**, 275-297.
[11] Ferreira, J. A. and Zwinderman A. H. (2006). On the Benjamini-Hochberg Method. *Annals of Statistics* **34**, 1827-1849.
[12] Gavrilov, Y., Benjamini, Y. and Sarkar, S. K. (2009). An adaptive step-down procedures with proven FDR control under independence. *Ann. Statist.* **37**, 619–629.
[13] Guo, W. (2009). A note on adaptive Bonferroni and Holm procedures under dependence. *Biometrika* **96**, 1012-1018.
[14] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple testing of significance. *Biometrika* **75**, 800–802.
[15] Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statist. Med.* **9**, 811–818.
[16] Holm, S. (1979). A simple sequentially rejective multiple testing procedure. *Scand. J. Statist.* **6**, 65–70.
[17] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383-386.
[18] Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655-660.
[19] Sarkar, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *Ann. Statist.* **26**, 494-504.
[20] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30**, 239–257.
[21] Sarkar, S. K. (2006). False discovery and false non-discovery rates in single-step multiple testing procedures. *Ann. Statist.* **34**, 394–415.
[22] Sarkar, S. K. (2008). On methods controlling the false discovery rate (with discussion). *Sankhya Ser A.* **70**, 135-168.
[23] Sarkar, S. K. and Chang, C-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc.*, **92**, 1601-1608.
[24] Sarkar, S. K. and Guo, W. (2009). On a generalized false discovery rate. *Ann. Statist.*, **37**, 337-363.
[25] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751-754.
[26] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. B* **66**, 187-205.
[27] Wu, W. (2008). On false discovery control under dependence. *Ann. Statist.* **36**, 364–380.

Index