ON STEPWISE CONTROL OF DIRECTIONAL ERRORS UNDER INDEPENDENCE AND SOME DEPENDENCE

By

Wenge Guo
Joseph P. Romano

Technical Report No. 2014-07
May 2014

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

ON STEPWISE CONTROL OF DIRECTIONAL ERRORS UNDER INDEPENDENCE AND SOME DEPENDENCE

By

Wenge Guo
New Jersey Institute of Technology

Joseph P. Romano
Stanford University

Technical Report No. 2014-07
May 2014

**This research was supported in part by
National Science Foundation grant DMS 0707085.**

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

ON STEPWISE CONTROL OF DIRECTIONAL ERRORS UNDER INDEPENDENCE AND SOME DEPENDENCE

Wenge Guo and Joseph P. Romano

New Jersey Institute of Technology and Stanford University

Abstract: In this paper, the problem of error control of stepwise multiple testing procedures is considered. For two-sided hypotheses, control of both type 1 and type 3 (or directional) errors is required, and thus mixed directional familywise error rate control and mixed directional false discovery rate control are each considered by incorporating both types of errors in the error rate. Mixed directional familywise error rate control of stepwise methods in multiple testing has proven to be a challenging problem, as demonstrated in Shaffer (1980). By an appropriate formulation of the problem, some new stepwise procedures are developed that control type 1 and directional errors under independence and various dependencies.

Key words and phrases: Directional error, false discovery rate, familywise error rate, Holm procedure, Hochberg procedure, multiple testing, positive dependence, stepwise procedure.

1 Introduction

The main problem considered in this paper is the construction of procedures for the simultaneous testing of n parameters θ_i. For convenience, the null hypotheses $\theta_i = 0$ are of interest. Of course, we would like to reject any null hypothesis if the data suitably dictates, but we also wish to make directional inferences about the signs of θ_i. First, consider the problem of simultaneously testing n null hypotheses against two-sided alternatives:

$$\check{H}_i : \theta_i = 0 \;\; vs. \;\; \check{H}_i' : \theta_i \neq 0, \quad i = 1, \ldots, n \; . \tag{1}$$

Suppose, for $i = 1, \ldots, n$, a test statistic T_i, is available for testing $\check{H}_i$. If $\check{H}_i$ is rejected, the decision regarding $\theta_i > 0$ (or $\theta_i < 0$) is made by checking if $T_i > 0$ (or $T_i < 0$). In making such rejection and directional decisions, three types of errors might occur. The first one is the usual type 1 error, which occurs when $\theta_i = 0$, but we falsely reject $\check{H}_i$ and declare $\theta_i \neq 0$. The second one is the type 2 error, which occurs when $\theta_i \neq 0$, but we fail to reject $\check{H}_i$. The last one is called type 3 or directional error, which occurs when $\theta_i > 0$ (or $\theta_i < 0$), but we falsely declare $\theta_i < 0$ (or $\theta_i > 0$). We wish to control both type 1 and type 3 errors at pre-specified levels and, subject to their control, find testing methods with small probability of type 2 errors.

Given any procedure which makes rejections as well as directional claims about any rejected hypotheses, let $\check{V}$ and $\check{S}$ denote the numbers of type 1 errors and type 3 errors, respectively, among $\check{R}$ rejected hypotheses. Let $\check{U} = \check{V} + \check{S}$ denoting the total number of type 1 and type 3 errors. Then, the usual familywise error rate (FWER) and false discovery rate (FDR) are defined respectively by FWER $= \Pr(\check{V} \geq 1)$ and FDR $= E\left(\check{V}/\max(\check{R}, 1)\right)$, and the mixed directional FWER and FDR are defined respectively by mdFWER $= \Pr(\check{U} \geq 1)$ and mdFDR $= E\left(\check{U}/\max(\check{R}, 1)\right)$.

The main objective of this paper is to develop stepwise procedures (described shortly) for controlling the mdFWER and mdFDR when simultaneously testing the n two-sided hypotheses $\check{H}_1, \ldots, \check{H}_n$. In multiple testing, the problem of simultaneously testing n two-sided hypotheses along with directional decisions subject to the control of the mdFWER is technically very challenge. Until now, only a few results have been obtained under the strong assumption of independence of the test statistics along with some additional conditions on the marginal distribution of the test statistics.

Shaffer (1980) proved that if the test statistics $T_i, i = 1, \ldots, n$ are mutually independent and if the distributions of the T_i's satisfy some additional conditions, the mdFWER of a directional Holm procedure is strongly controlled at level α. She also constructed a counterexample where the aforementioned procedure loses the control of the mdFWER even under independence when the test statistics are Cauchy distributed. Holm (1979b, 1981) extended Shaffer's (1980)'s result to normal distributional settings where the T_i's are conditionally independent. Finner (1994) and Liu (1997) independently used Shaffer's (1980) method of proof to show the mdFWER control of directional Hochberg procedure by making the same distributional assumptions as Shaffer (1980). By generalizing Shaffer's method of proof, Finner (1999) extended Shaffer's result on the Holm procedure to a large class of stepwise or closed multiple testing procedures under the same assumptions as in Shaffer (1980). He also gave a new but very simple and elegant proof for the aforementioned result under the assumption of TP_3 densities. For further discussions on the mdFWER control of closed testing methods, see Westfall, Bretz and Tobias (2013).

Another method to tackle the problem of directional errors has been considered in Bauer, Hackle, Hommel and Sonnemann (1986), in which the problem of testing n two-sided hypotheses testing with additional directional decisions are reformulated as the problem of testing n pairs of one-sided hypotheses given by

$$H_{i1} : \theta_i \leq 0 \ \text{ vs. } \ H_{i1}^{'} : \theta_i > 0 \ ,$$

and

$$\tilde{H}_{i2} : \theta_i \geq 0 \ \text{ vs. } \ \tilde{H}_{i2}^{'} : \theta_i < 0$$

for $i = 1, \ldots, n$. They proved that without additional distributional assumptions, only a slight improvement of the conventional Holm procedure is possible for testing these $2n$ hypotheses. They also showed by a counterexample that in general distributional settings, a further improvement of their procedure is impossible. Compared with Shaffer's (1980) directional Holm procedure for testing n two-sided hypotheses, their procedure is very conservative, although it controls directional errors under more general distributional settings of arbitrary dependence.

Finally, they also reformulated the aforementioned problem as the problem of testing n pairs of one-sided hypotheses given by

$$H_{i1} : \theta_i \leq 0 \ \text{ vs. } \ H_{i1}^{'} : \theta_i > 0 \ ,$$

and

$$H_{i2} : \theta_i > 0 \ \text{ vs. } \ H_{i2}^{'} : \theta_i < 0 \ ,$$

for $i = 1, \ldots, n$, among which there is exactly one true null hypothesis within each pair of one-sided hypotheses. They proved that the modified Bonferroni procedure with the critical constant α/n (as opposed to $\alpha/2n$) strongly controls the FWER when testing these $2n$ one-sided hypotheses. This result is of course trivial because in this formulation there are exactly n true null hypotheses. At the same time, given that there are always n true null hypotheses, it is perhaps surprising that one can, as we do, develop stepdown methods that improve upon this single step method. (Indeed, at any step when applying a stepdown method, there are always n true null hypotheses, and this number does not reduce.)

In the above two formulations of one-sided hypotheses, there are some inherent disadvantages when developing stepwise methods for controlling the FWER. In the first formulation, there may be a different number of true null hypotheses between $\theta_i = 0$ and $\theta_i \neq 0$, which makes it challenging to develop powerful stepwise methods in this formulation, as shown in Bauer et al. (1986). In the second formulation, one possible type 1 error will not be counted even though T_i is very small when $\theta_i = 0$, which makes it unable to completely control type 1 and type 3 errors in the original formulation of two-sided hypotheses even though the FWER is controlled in this formulation. Further discussion of this point will be presented later. On the other hand, the problem of the mdFDR control seems to be technically less challenge and methods for controlling the mdFDR are available (see Benjamini and Yekutieli, 2005; Guo, Sarkar and Peddada, 2010).

In the next section, some basic notation is given, as well as our approach to the problem. Theorems 1–4 deal with control of the familywise error rate with directional decisions, first under independence, and then under block dependence and positive dependence. Theorems 5–8 analogously provide results for the false discovery rate.

2 Preliminaries

In this section, some necessary notation and basic concepts are introduced.

2.1 Notation

Suppose T_i has cumulative distribution function $F_{i,\theta_i}(\cdot)$ (with density denoted $f_{i,\theta_i}(\cdot)$ when it is well-defined), both of which depend on a single parameter θ_i. It is assumed that the null distribution of T_i, i.e. $F_{i,0}(\cdot)$ is continuous. We also assume that $F_{i,\theta_i}(t)$ is non-increasing in θ_i for any given t and $F_{i,0}(t)$ is symmetric about zero, i.e., $F_{i,0}(-t) = 1 - F_{i,0}(t)$ for any t. (In fact, the symmetry assumption is not really necessary; indeed, one may take the probability integral transformation $F_{i,0}(T_i)$ to get a new test statistic that is uniform and then shift it by 1/2 to get a "symmetric" null test statistic.) Let t_i be the observed value of T_i. Then, the (two-sided) p-value for testing $\check{H}_i$ is

$$\check{P}_i = 2 \min(F_{i,0}(t_i), 1 - F_{i,0}(t_i)) \ .$$

Let $\check{P}_{(1)} \leq \ldots \leq \check{P}_{(n)}$ be the ordered p-values and $\check{H}_{(1)}, \ldots, \check{H}_{(n)}$ the associated null hypotheses. Then, given a non-decreasing set of critical constants $0 < \alpha_1 \leq \cdots \leq \alpha_n < 1$, a stepdown multiple testing procedure rejects the set of null hypotheses $\{\check{H}_{(i)}, i \leq i^*_{SD}\}$ and accepts the rest, where $i^*_{SD} = \max\{i : \check{P}_{(j)} \leq \alpha_j \ \forall \ j \leq i\}$ if the maximum exists, and otherwise it accepts all the null hypotheses. A stepup procedure, on the other hand, rejects the set $\{\check{H}_{(i)}, i \leq i^*_{SU}\}$ and accepts the rest, where $i^*_{SU} = \max\{i : \check{P}_{(i)} \leq \alpha_i\}$ if the maximum exists, otherwise it accepts all the null hypotheses. Furthermore, if stepwise procedures (stepdown or stepup) are applied along with additional directional decisions, such procedures are often termed as directional stepwise procedures (Shaffer, 2002). (A stepwise procedure with constant α_i is referred to as a single-step procedure.) The constants in a stepwise procedure are determined subject to the control of a suitable error rate at a pre-specified level α.

2.2 Formulation

In order to further explore the problem of controlling type 1 and type 3 errors under independence, and also under some dependence, we first reformulate this problem as an equivalent one of simultaneously testing multiple one-sided hypotheses subject to the control of the FWER (or FDR). Specifically, $\check{H}_i, i = 1, \ldots, n$ is partitioned as three one-sided hypotheses,

$$H_{i1} : \theta_i \leq 0 \ \text{ vs. } \ H'_{i1} : \theta_i > 0 \ ,$$

$$H_{i2} : \theta_i > 0 \ \text{ vs. } \ H'_{i2} : \theta_i < 0 \ ,$$

and

$$H_{i3} : \theta_i = 0 \ \text{ vs. } \ H'_{i3} : \theta_i < 0 \ .$$

As we know, for the original problem of testing the two-sided hypotheses $\check{H}_i, i = 1, \ldots, n$ along with directional decisions, there are two possibilities of type 1 errors and two possibilities of type 3 errors. Indeed, when $\theta_i = 0$, the corresponding test statistic T_i can be too large or too small; or, when $\theta_i > 0$ (or < 0), T_i is too small (or large). In the new formulation, those two possible directional errors in the original problem are transformed as type 1 errors for testing H_{i1} and H_{i2}, respectively, and the two possible type 3 errors when testing $\check{H}_i$ are transformed as type 1 errors for testing H_{i1} and H_{i3}, respectively. It should be noted that the additional directional decisions in all these formulations of one-sided hypotheses is unnecessary as any rejection already corresponds to a directional decision. Note that, when T_i is used for testing H_{i1}, $-T_i$ is used for testing both H_{i2} and H_{i3}.

Let $\mathcal{F} = \{H_{ij} : i = 1, \ldots, n, j = 1, 2, 3\}$ denote the whole family of the $3n$ one-sided hypotheses H_{ij}'s to be tested. We split $\mathcal{F}$ as two subfamilies $\mathcal{F}_1$ and $\mathcal{F}_2$, where

$$\mathcal{F}_1 = \{H_{ij} : i = 1, \ldots, n, j = 1, 2\}$$

and

$$\mathcal{F}_2 = \{H_{i3} : i = 1, \ldots, n\} .$$

In this paper, we use a separate approach for testing multiple families of hypotheses. In this approach, two given multiple testing methods are used for testing $\mathcal{F}_i, i = 1, 2$, respectively. If, let $S_i, i = 1, 2$ denote the respective rejection sets of testing $\mathcal{F}_i$, then the rejection set for testing $\mathcal{F}$ is $S_1 \bigcup S_2$. The advantage of splitting $\mathcal{F}$ derives from the fact that $\mathcal{F}_1$ consists of $2n$ one-sided hypotheses, of which exactly n of them correspond to true null hypotheses.

For the aforementioned approach, let V_i denote the number of type 1 errors among R_i rejected hypotheses when testing $\mathcal{F}_i$ for $i = 1, 2$, and let V denote the number of type 1 errors among R rejected hypotheses when testing $\mathcal{F}$. Thus, $R = R_1 + R_2$ and $V = V_1 + V_2$. Then, the FWER and FDR of the multiple testing method for testing $\mathcal{F}$ are defined respectively by

$$\text{FWER}_{\mathcal{F}} = Pr\{V \geq 1\}$$

and

$$\mathrm{FDR}_{\mathcal{F}} = E\left(V/\max(R,1)\right) \ .$$

Similarly, the FWER and FDR for testing $\mathcal{F}_i$ are defined respectively by $\mathrm{FWER}_{\mathcal{F}_i} = \Pr\{V_i \geq 1\}$ and $\mathrm{FDR}_{\mathcal{F}_i} = E\left(V_i/\max(R_i,1)\right), i = 1,2$. Note that $V \geq 1$ implies $V_1 \geq 1$ or $V_2 \geq 1$, so that $\mathrm{FWER}_{\mathcal{F}} \leq \mathrm{FWER}_{\mathcal{F}_1} + \mathrm{FWER}_{\mathcal{F}_2}$. Similarly, using the simple inequality

$$\frac{V}{\max(R,1)} = \frac{V_1}{\max(R,1)} + \frac{V_2}{\max(R,1)} \leq \frac{V_1}{\max(R_1,1)} + \frac{V_2}{\max(R_2,1)} \ ,$$

we have $\mathrm{FDR}_{\mathcal{F}} \leq \mathrm{FDR}_{\mathcal{F}_1} + \mathrm{FDR}_{\mathcal{F}_2}$. We will develop in this paper respective stepwise methods for controlling the $\mathrm{FWER}_{\mathcal{F}}$ and $\mathrm{FDR}_{\mathcal{F}}$ when testing $\mathcal{F}$ based on the aforementioned separate approach and the above two inequalities. We note that in the existing literature, a number of powerful stepwise methods have been introduced under various dependencies for testing $\mathcal{F}_2$, for which unlike $\mathcal{F}_1$, there is no specific dependency relationship among the test statistics corresponding to those one-sided hypotheses in $\mathcal{F}_2$. For example, control of the $\mathrm{FWER}_{\mathcal{F}_2}$ can be done by the Holm (1979a) and Hochberg (1988) while the Benjamini and Hochberg (1995) procedure (BH) can be used to control the $\mathrm{FDR}_{\mathcal{F}_2}$ (Therefore, through much of the paper, we will focus on developing stepwise methods for controlling the $\mathrm{FWER}_{\mathcal{F}_1}$ and $\mathrm{FDR}_{\mathcal{F}_1}$ under independence and certain dependencies, unless noted otherwise.

Before we embark upon control of any error rate for $\mathcal{F}_1$ as a building block for control over the larger family $\mathcal{F}$, we would like to argue that this seemingly more restrictive control over the smaller family $\mathcal{F}_1$ is already a plausible approach to the problem of control of directional errors. For this, we draw upon the wisdom and philosophy of one of the fathers in the field of multiple testing, John Tukey. In the context of single testing, Tukey argued that a point null hypothesis is never true, and therefore control of type 1 errors is the wrong formulation. Tukey cared more about whether or not one could tell the "effect size" or the "sign" of a parameter. To quote Tukey (1991), "Statisticians classically asked the wrong question – and were willing to answer with a lie, one that was often a downright lie.......All we know about the world teaches us that the effects of A and B are always different – in some decimal place – for any A and B. Thus asking 'Are the effects different' is foolish. What we should be answering first is 'Can we tell the direction in which the effects of A differ from the effects of B?'. " Thus, for Tukey, emphasis must be completely upon control of directional or type 3 errors. So, as also argued in Jones and Tukey (2000) in the context of a test of a single parameter θ (which is motivated there as a difference in means), one can and should apply a classical two-sided t-test so that the probability of observing an outcome in either the right or left tail is not $\alpha/2$, but α. That is, if one wishes to make directional inferences or claims about a parameter (which is always desirable) then the problem of testing the null hypothesis $\theta = 0$ at level α should be replaced by the problem of testing the two hypotheses: testing $\theta < 0$ against $\theta > 0$ as well as testing $\theta > 0$ against $\theta < 0$. Since $\theta = 0$ never holds, one can always use the $1 - \alpha$ quantile in the right tail rather than the $1 - \alpha/2$ quantile, and similarly the α quantile in the left tail. In our context, if we acknowledge that θ_i is never 0 from the start, then we never need to include $\mathcal{F}_2$ in the family of hypothesis tested, and the problem of control of directional errors is equivalent to control of the error rate over $\mathcal{F}_1$. Moreover, if one takes Tukey's stance to heart, then the inequality in the definition of H_{i1} can be a strict inequality. However, we retain the inequality because the methods we develop apply to H_{i1} as defined, and hence to the more restricted definition. Thus, control over $\mathcal{F}_1$ is emphasized throughout, as both a building block toward control over $\mathcal{F}$ but also as a formulation worth studying in its own right. A nice review of Tukey's contributions to multiple testing can be found in Benjamini and Braun (2002).

2.3 Assumptions

It should be noted that $H_{i1} \bigcap H_{i2}$ is empty and $H_{i1} \bigcup H_{i2}$ is the whole parameter space. Thus, there are exactly n true and n false null hypotheses in $\mathcal{F}_1$. For notational convenience, we respectively use $H_1, \ldots, H_n$ and $H_{n+1}, \ldots, H_{2n}$ denoting the n true and n false nulls, (H_i, H_{n+i}) denoting the pair of true and false nulls (H_{i1}, H_{i2}), and (P_i, P_{n+i}) denoting the pair of the corresponding (one-sided) p-values. With the test statistic T_i and the calculated value t_i, the p-value P_i corresponding to H_i is equal to $F_{i,0}(t_i)$ or $1 - F_{i,0}(t_i)$ depending on $H_i : \theta_i \leq 0$ or $H_i : \theta_i > 0$, and $P_{n+i} = 1 - P_i$ for $i = 1, \ldots, n$. In addition, let $I_0 = \{1, \ldots, n\}$ and $I_1 = \{n+1, \ldots, 2n\}$ denote the index sets of true and false nulls among the $2n$ hypotheses, $H_1, \ldots, H_{2n}$, respectively.

Regarding the marginal distribution of the true null p-values, the following assumptions are invoked throughout much of the paper:

A.1 For any p-value $P_i, i \in I_0$ and true parameter θ_i,

$$\Pr_{\theta_i} \{P_i \leq p\} \leq p \quad \text{for any} \quad 0 \leq p \leq 1. \tag{2}$$

For $\theta_i = 0$, (2) is an equality; that is, $P_i \sim U(0,1)$ for $i \in I_0$ when $\theta_i = 0$.

A.2 For any p-value $P_i, i \in I_0$ and given parameter θ_i,

$$\Pr_{\theta_i} \left\{ P_i \leq p \middle| P_i \leq p' \right\} \leq \Pr_{\theta_i = 0} \left\{ P_i \leq p \middle| P_i \leq p' \right\} \; . \tag{3}$$

for any $0 \leq p \leq p' \leq 1$.

A.3 The test statistics $T_i, i = 1, \ldots, n$ are mutually independent.

While the assumption of independence is quite restrictive, to the best of our knowledge, all the previous results on the mdFWER control of the existing stepwise procedures along with directional decisions are established under this assumption. However, not all of our results require both A2 and A3.

Of course, under assumption A.1, the right hand side of (3) is just p/p'. Assumption A.2 is easily satisfied by the usual test statistics. Actually, the following result holds.

Lemma 2.1 *If the family of densities $f_{i,\theta}(\cdot)$ of T_i satisfies the assumption of monotone* *likelihood ratio (MLR), i.e., for any given $\theta_1 > \theta_0$ and $x_1 > x_0$, $\frac{f_{i,\theta_1}(x_1)}{f_{i,\theta_0}(x_1)} \geq \frac{f_{i,\theta_1}(x_0)}{f_{i,\theta_0}(x_0)}$, then Assumption A.2 holds.*

For the proof of Lemma 1, see the Appendix. Of course, the assumption holds if the distribution of T_i is a normal shift model, which often asymptotically approximates the underlying situation.

By Lemma 2.1, the MLR assumption implies Assumption A.2. However, these two assumptions are not equivalent. Assumption A.2 is slightly weaker than the MLR assumption. It is equivalent to the following condition: for any given θ_1 and $x_1 > x_0$, $\frac{F_{i,\theta_1}(x_1)}{F_{i,0}(x_1)} \geq \frac{F_{i,\theta_1}(x_0)}{F_{i,0}(x_0)}$ when $\theta_1 > 0$ and $\frac{1-F_{i,\theta_1}(x_1)}{1-F_{i,0}(x_1)} \leq \frac{1-F_{i,\theta_1}(x_0)}{1-F_{i,0}(x_0)}$ when $\theta_1 < 0$. It should be pointed out that Assumption A.2 is different from the conventional TP_2-property of $\partial[1 - F_{i,\theta_i}(x)]/\partial\theta_i$, which is almost always assumed in the existing literature on control of directional errors (Shaffer, 1980; Finner, 1999). The only exception is Sarkar, Sen and Finner (2004). In that paper, it is assumed that $f_{i,\theta_i}(\cdot)$ satisfies the aforementioned MLR condition.

To characterize the joint distribution among the test statistics $T_i, i = 1, \ldots, n$, several dependence assumptions have been made in this paper: independence, within-block dependence, between-block dependence, and positive dependence. The positive dependence condition, which will be of the type characterized by the following:

$$E\left\{\phi(T_1, \ldots, T_n) \mid T_i \geq u\right\} \uparrow u \in (0, 1), \tag{4}$$

for each T_i and any (coordinatewise) non-decreasing function ϕ. This type of positive dependence is commonly encountered and used in multiple testing; see, for instance, Sarkar (2008) for references. Other dependence conditions such as independence, within-block and between-block dependence, will be characterized in Sections 3 and 4, respectively.

Finally, for convenience of discussion, the following notation will be used. Given any index set of false null hypotheses, $S \subset I_1$, define $\overline{S} = I_1 \backslash S$, $S_{-n} = \{i \in I_0 : n + i \in S\}$, and $\overline{S}_{-n} = \{i \in I_0 : n + i \in \overline{S}\}$. It is easy to see that $|S| = |S_{-n}|$ and $|\overline{S}| = |\overline{S}_{-n}|$.

3 Controlling the mdFWER under independence

In this section, several stepwise procedures for controlling the $\text{FWER}_{\mathcal{F}_1}$ are presented under the assumption of independence.

3.1 Two-stage procedure

For simplicity, we first consider a two-stage version of the usual Holm procedure for testing $\mathcal{F}_1$ as follows.

- Stage 1. Reject all null hypotheses H_i with the p-values less than or equal to α/n. Let r be the total number of rejections at stage 1. If $r = n$, we stop testing; otherwise,
- Stage 2. For the remaining hypotheses, reject those with the p-values less than or equal to $\alpha/(n - r)$.

In the above two-stage procedure, the Bonferroni procedure is used in the first stage for testing the $2n$ hypotheses. Generally, the Bonferroni would actually use the critical constant $\alpha/2n$ when testing $\mathcal{F}_1$. However, in this formulation we know there are exactly n true null hypotheses in $\mathcal{F}_1$ and we can apply an obviously modified Bonferroni procedure with critical constant α/n. Our method then improves upon this with a second stage improvement in the spirit of a stepdown method. The two-stage procedure can also be regarded as an adaptive Bonferroni procedure with the critical constant $c = \alpha/\max(n - R_1(\alpha/n), 1)$, where $R_1(\alpha/n) = \sum_{i=1}^{2n} I(P_i \leq \alpha/n)$ (Finner and Gontscharuk, 2009; Guo, 2009).

Let R_{11} be the index set of rejected false null hypotheses at the first stage, R_{10} be the index set of true null hypotheses for which the corresponding p-values less than $1-\alpha/n$, and $R_{10}^{(-j)}$ be the index set of true null hypotheses excluding H_j for which the corresponding p-

values less than $1-\alpha/n$, that is, $R_{11} = \{i \in I_1 : P_i \leq \alpha/n\}, R_{10} = \{i \in I_0 : P_i < 1-\alpha/n\}$, and $R_{10}^{(-j)} = \{i \in I_0 \backslash \{j\} : P_i < 1-\alpha/n\} = R_{10} \backslash \{j\}$.

Let $\widehat{P}_{(1)}^{I_0}$ be the minimum p-value corresponding to the true null hypotheses with indices in I_0, for any given parameter vector $\theta = (\theta_1, \ldots, \theta_n)$, we have

$$
\begin{aligned}
& \mathrm{FWER}_{\mathcal{F}_1}(\theta) = \sum_{S \subset I_1} \Pr_\theta \left\{ R_{11} = S, \widehat{P}_{(1)}^{I_0} \leq \frac{\alpha}{n-|S|} \right\} \\
= & \sum_{S \subset I_1} \Pr_\theta \left\{ P_i \leq \frac{\alpha}{n} \text{ for all } i \in S, P_i > \frac{\alpha}{n} \text{ for all } i \in \overline{S}, \widehat{P}_{(1)}^{I_0} \leq \frac{\alpha}{n-|S|} \right\} \\
= & \sum_{S_{-n} \subset I_0} \Pr_\theta \left\{ P_i \geq 1-\frac{\alpha}{n} \text{ for all } i \in S_{-n}, P_i < 1-\frac{\alpha}{n} \text{ for all } i \in \overline{S}_{-n}, \widehat{P}_{(1)}^{\overline{S}_{-n}} \leq \frac{\alpha}{n-|S_{-n}|} \right\} \\
\leq & \sum_{S_{-n} \subset I_0} \sum_{j \in \overline{S}_{-n}} \Pr_\theta \left\{ P_i \geq 1-\frac{\alpha}{n} \text{ for all } i \in S_{-n}, P_i < 1-\frac{\alpha}{n} \text{ for all } i \in \overline{S}_{-n}, P_j \leq \frac{\alpha}{n-|S_{-n}|} \right\} \\
= & \sum_{S_{-n} \subset I_0} \sum_{j \in \overline{S}_{-n}} \Pr_\theta \left\{ R_{10}^{(-j)} = \overline{S}_{-n} \backslash \{j\}, P_j \leq \frac{\alpha}{n-|S_{-n}|} \right\}. \qquad (5)
\end{aligned}
$$

The inequality follows from the Bonferroni inequality.

Note that for $j \in \overline{S}_{-n}$,

$$
\begin{aligned}
& \Pr_\theta \left\{ R_{10}^{(-j)} = \overline{S}_{-n} \backslash \{j\}, P_j \leq \frac{\alpha}{n-|S_{-n}|} \right\} \\
= & \Pr_\theta \left\{ R_{10}^{(-j)} = \overline{S}_{-n} \backslash \{j\}, P_j \leq \frac{\alpha}{n-|S_{-n}|}, P_j \leq 1-\frac{\alpha}{n} \right\} \\
= & \Pr_\theta \left\{ R_{10}^{(-j)} = \overline{S}_{-n} \backslash \{j\}, P_j \leq \frac{\alpha}{n-|S_{-n}|} \middle| P_j \leq 1-\frac{\alpha}{n} \right\} \Pr_{\theta_j} \left\{ P_j \leq 1-\frac{\alpha}{n} \right\} \\
= & \Pr_{\theta^{(-j)}} \left\{ R_{10}^{(-j)} = \overline{S}_{-n} \backslash \{j\} \right\} \Pr_{\theta_j} \left\{ P_j \leq 1-\frac{\alpha}{n} \right\} \Pr_{\theta_j} \left\{ P_j \leq \frac{\alpha}{n-|S_{-n}|} \middle| P_j \leq 1-\frac{\alpha}{n} \right\} \\
\leq & \Pr_\theta \left\{ R_{10} = \overline{S}_{-n} \right\} \Pr_{\theta_j=0} \left\{ P_j \leq \frac{\alpha}{n-|S_{-n}|} \middle| P_j \leq 1-\frac{\alpha}{n} \right\} \\
= & \Pr_\theta \left\{ R_{10} = \overline{S}_{-n} \right\} \frac{\alpha}{n-|S_{-n}|} \frac{1}{1-\alpha/n}, \qquad (6)
\end{aligned}
$$

where $\theta^{(-j)} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_n)$. Here, the third equality follows from assumption A.3 and the fourth follows from assumption A.1 under which $P_j \sim U(0,1)$ when $\theta_j = 0$. For the inequality, the first term of its right-hand side follows from assumption A.3 under which the first two terms of the left-hand side match up, and the second one of its right-hand side follows from assumption A.2.

Applying (6) to (5), we have

$$\begin{aligned}
\mathrm{FWER}_{\mathcal{F}_1}(\theta) &\le \sum_{S_{-n}\subset I_0}\sum_{j\in\overline{S}_{-n}} \Pr_\theta\left\{R_{10}=\overline{S}_{-n}\right\}\frac{\alpha}{n-|S_{-n}|}\frac{1}{1-\alpha/n}\\
&= \sum_{S_{-n}\subset I_0} \Pr_\theta\left\{R_{10}=\overline{S}_{-n}\right\}\frac{\alpha}{1-\alpha/n}\\
&\le \frac{\alpha}{1-\alpha/n}. \qquad (7)
\end{aligned}$$

Theorem 3.1 *Consider the two-stage procedure defined as above. Under assumption A.1 - A.3, the following conclusions hold.*

(i) *The procedure strongly controls the $FWER_{\mathcal{F}_1}$ at level $\frac{\alpha}{1-\alpha/n}$.*

(ii) $\limsup_{n\to\infty} FWER_{\mathcal{F}_1} \le \alpha$. *That is, the procedure asymptotically controls the $FWER_{\mathcal{F}_1}$ at level α. Moreover, if the critical constants of the two-stage procedure are rescaled by using $\frac{\alpha}{1+\alpha/n}$ to replace α, then the resulting procedure strongly controls the $FWER_{\mathcal{F}_1}$ at level α even in finite samples.*

Remark 3.1 The upper bound of the $\mathrm{FWER}_{\mathcal{F}_1}$ in (7) is only slightly larger than α. For example, as $\alpha = 0.05$ and $n = 10$, the upper bound is about 0.05025. In addition, we know that $\frac{1}{1-x} \le 1 + x + 2x^2$ when $0 \le x \le 0.5$. Thus,

$$\frac{\alpha}{1-\alpha/n} \le \alpha + \alpha^2/n + 2\alpha^3/n^2. \qquad (8)$$

From (8), we can see how close the upper bound is to the pre-specified level α.

Remark 3.2 It should be noted that the above rescaled two-stage procedure in Theorem 3.1 is not consistently more powerful than Bauer et al. (1986)'s modified Bonferroni procedure with the critical constant α/n, since its critical constant at stage 1 is slightly smaller than α/n. However, by carefully checking the proof of Theorem 3.1 (ii), we can see that for the original two-stage procedure, we actually only need to rescale its critical constant at stage 2 in order to maintain the control of the FWER at level α. The newly modified procedure is described in details as follows.

- Stage 1. Reject all null hypotheses H_i with the p-values less than or equal to α/n. Let r be the total number of rejections at stage 1. If $r = n$, we stop testing; otherwise,
- Stage 2. For the remaining hypotheses, reject those with the p-values less than or equal to $\beta/(n-r)$, where $\beta = \frac{\alpha}{1+\alpha/n}$.

It is easy to see that the above procedure is consistently more powerful than Bauer et al.'s modified Bonferroni procedure, because for this procedure, even though only one hypothesis is rejected at stage 1, its critical constant $\frac{\alpha}{(n-1)(1+\alpha/n)}$ at stage 2 is also larger than α/n, the critical constant of Bauer et al.'s procedure.

Although the upper bound of the $\text{FWER}_{\mathcal{F}_1}$ of the original two-stage procedure is only slightly larger than α, this procedure cannot always control the $\text{FWER}_{\mathcal{F}_1}$ at level α in the finite samples. In the following, we present an example where the FWER of the aforementioned procedure when testing $\mathcal{F}_1$ is above α but of course below $\alpha/(1-\alpha/n)$ as proved in Theorem 3.1.

Example 1 Consider the special case of $\theta = (\theta_1, \dots, \theta_n) \to 0$, thus $P_i \sim U(0,1)$ for all

$i \in I_0$. By using the third equality of (5), we have

$$\begin{aligned}
&\text{FWER}_{\mathcal{F}_1}(\theta) \\
= &\sum_{r=0}^{n-1} \sum_{\substack{S_{-n} \subset I_0 \\ |S_{-n}|=r}} \Pr_\theta \left\{ P_i \geq 1 - \frac{\alpha}{n} \text{ for } i \in S_{-n}, P_i < 1 - \frac{\alpha}{n} \text{ for } i \in \overline{S}_{-n}, \widehat{P}_{(1)}^{\overline{S}_{-n}} \leq \frac{\alpha}{n-r} \right\} \\
= &\sum_{r=0}^{n-1} \sum_{\substack{S_{-n} \subset I_0 \\ |S_{-n}|=r}} \Pr_\theta \left\{ P_i \geq 1 - \frac{\alpha}{n} \text{ for } i \in S_{-n} \right\} \Pr_\theta \left\{ P_i < 1 - \frac{\alpha}{n} \text{ for } i \in \overline{S}_{-n}, \widehat{P}_{(1)}^{\overline{S}_{-n}} \leq \frac{\alpha}{n-r} \right\} \\
= &\sum_{r=0}^{n-1} \sum_{\substack{S_{-n} \subset I_0 \\ |S_{-n}|=r}} \left(\frac{\alpha}{n}\right)^r \Pr_\theta \left\{ P_i < 1 - \frac{\alpha}{n} \text{ for } i \in \overline{S}_{-n}, \widehat{P}_{(1)}^{\overline{S}_{-n}} \leq \frac{\alpha}{n-r} \right\}. \qquad (9)
\end{aligned}$$

In the above special case with $|S_{-n}| = r$, we have

$$\begin{aligned}
&\Pr_\theta \left\{ P_i < 1 - \frac{\alpha}{n} \text{ for all } i \in \overline{S}_{-n}, \widehat{P}_{(1)}^{\overline{S}_{-n}} \leq \frac{\alpha}{n-r} \right\} \\
= &\Pr_\theta \left\{ \widehat{P}_{(1)}^{\overline{S}_{-n}} \leq \frac{\alpha}{n-r} \big| P_i < 1 - \frac{\alpha}{n} \text{ for all } i \in \overline{S}_{-n} \right\} \Pr_\theta \left\{ P_i < 1 - \frac{\alpha}{n} \text{ for all } i \in \overline{S}_{-n} \right\} \\
= &\left[1 - \Pr_\theta \left\{ \widehat{P}_{(1)}^{\overline{S}_{-n}} > \frac{\alpha}{n-r} \big| P_i < 1 - \frac{\alpha}{n} \text{ for all } i \in \overline{S}_{-n} \right\} \right] \prod_{i \in \overline{S}_{-n}} \Pr_\theta \left\{ P_i < 1 - \frac{\alpha}{n} \right\} \\
= &\left[1 - \prod_{i \in \overline{S}_{-n}} \Pr_\theta \left\{ P_i > \frac{\alpha}{n-r} \big| P_i < 1 - \frac{\alpha}{n} \right\} \right] \left(1 - \frac{\alpha}{n}\right)^{n-r} \\
= &\left[1 - \left(1 - \frac{\frac{\alpha}{n-r}}{1 - \frac{\alpha}{n}} \right)^{n-r} \right] \left(1 - \frac{\alpha}{n}\right)^{n-r} \\
= &\left(1 - \frac{\alpha}{n}\right)^{n-r} - \left(1 - \frac{\alpha}{n} - \frac{\alpha}{n-r}\right)^{n-r}. \qquad (10)
\end{aligned}$$

The second and third equalities follow from assumption A.3. Apply (10) into (9), we have

$$\text{FWER}_{\mathcal{F}_1}(\theta) = \sum_{r=0}^{n-1} \binom{n}{r} \left(\frac{\alpha}{n}\right)^r \left[\left(1 - \frac{\alpha}{n}\right)^{n-r} - \left(1 - \frac{\alpha}{n} - \frac{\alpha}{n-r}\right)^{n-r} \right]. \qquad (11)$$

Through simple algebra calculation, we find out that $\text{FWER}_{\mathcal{F}_1}(\theta) = \alpha + \frac{\alpha^2}{4} > \alpha$ as $n = 2$ and $\text{FWER}_{\mathcal{F}_1}(\theta) = \alpha + \frac{\alpha^3}{108} > \alpha$ as $n = 3$. Thus, the original two-stage procedure and thereby the usual Holm procedure with the critical values $\alpha_i = \frac{\alpha}{n-i+1}, i = 1, \ldots, n$, cannot always control the $\text{FWER}_{\mathcal{F}_1}$ at level α. ■

It should be noted that in the above example, assumption A.2 is not used. This example shows that no matter whether or not assumption A.2 holds, the aforementioned two-stage procedure cannot control the FWER at level α in the finite samples.

3.2 Holm-type stepdown procedure

Consider a modified Holm procedure, which is a stepdown procedure with the first n critical values $\alpha_i = \frac{\alpha}{n-i+1+\alpha}, i = 1, \ldots, n$ for testing $\mathcal{F}_1$. Let $\widehat{q}_{(1)} \leq \ldots \leq \widehat{q}_{(n)}$ denote the ordered false null p-values. Define $J = \max\{j : \widehat{q}_{(i)} \leq \alpha_i, \ \forall i \leq j\}$, provided this maximum exists; otherwise, let $J = 0$. Let K denote the index set of the J rejected false null hypotheses when applying the stepdown procedure to simultaneously test the n false null hypotheses $H_{n+1}, \ldots, H_{2n}$, and E_1 denote the event of at least one falsely rejected hypothesis when applying the same procedure to simultaneously test $H_1, \ldots, H_{2n}$. It should be noted that if $J = n$, then no true null hypotheses are falsely rejected when testing $\mathcal{F}_1$. Thus,

$$\begin{aligned} E_1 &= \bigcup_{j=0}^{n-1} \left\{ J = j, \widehat{P}_{(1)}^{I_0} \leq \alpha_{j+1} \right\} \\ &= \bigcup_{j=0}^{n-1} \bigcup_{\substack{S \subset I_1 \\ |S| = j}} \left\{ K = S, \widehat{P}_{(1)}^{I_0} \leq \alpha_{j+1} \right\} \\ &= \bigcup_{S \subset I_1} \left\{ K = S, \widehat{P}_{(1)}^{\overline{S}_{-n}} \leq \alpha_{|S|+1} \right\}. \end{aligned} \tag{12}$$

Hence,

$$\begin{aligned}
\mathrm{FWER}_{\mathcal{F}_1}(\theta) &= \Pr_\theta(E_1) \\
&= \sum_{S\subset I_1} \Pr_\theta\left\{K=S, \widehat{P}_{(1)}^{\overline{S}_{-n}} \le \alpha_{|S|+1}\right\} \\
&\le \sum_{S\subset I_1}\sum_{j\in\overline{S}_{-n}} \Pr_\theta\left\{K=S, P_j \le \alpha_{|S|+1}\right\} \\
&= \sum_{S\subset I_1}\sum_{j\in\overline{S}} \Pr_\theta\left\{K^{\{-j\}}=S, P_j \ge 1-\alpha_{|S|+1}\right\}, \qquad (13)
\end{aligned}$$

where $K^{\{-j\}}$ is the index set of rejected false null hypotheses by using the stepdown procedure with the critical constants $\alpha_i = \frac{\alpha}{n-i+1+\alpha}, i = 1, \ldots, n-1$ to simultaneously test the $n-1$ false null hypotheses $H_{n+1}, \ldots, H_{2n}$ excluding H_j with $j \in I_1$.

By using the similar argument lines as in (6), we have

$$\begin{aligned}
&\Pr_\theta\left\{K^{\{-j\}}=S, P_j \ge 1-\alpha_{|S|+1}\right\} \\
=\ &\Pr_\theta\left\{K^{\{-j\}}=S, P_j > \alpha_{|S|+1}\right\} \Pr_{\theta_j}\left\{P_j \ge 1-\alpha_{|S|+1} \,\middle|\, P_j > \alpha_{|S|+1}\right\} \\
\le\ &\Pr_\theta\left\{K=S\right\} \Pr_{\theta_j=0}\left\{P_j \ge 1-\alpha_{|S|+1} \,\middle|\, P_j > \alpha_{|S|+1}\right\} \\
=\ &\Pr_\theta\left\{K=S\right\} \alpha_{|S|+1}\frac{1}{1-\alpha_{|S|+1}} \\
=\ &\Pr_\theta\left\{K=S\right\} \frac{\alpha}{n-|S|}. \qquad (14)
\end{aligned}$$

Applying (14) to (13), we have

$$\mathrm{FWER}_{\mathcal{F}_1}(\theta) \le \alpha \sum_{S\subset I_1} \Pr_\theta\left\{K=S\right\} \le \alpha. \qquad (15)$$

Therefore, the following conclusion holds.

Theorem 3.2 *Consider the stepdown procedure with the critical constants* $\alpha_i = \frac{\alpha}{n-i+1+\alpha}, i = 1, \ldots, n$. *Under assumption A.1 - A.3, the procedure strongly controls the* $FWER_{\mathcal{F}_1}$ *at level* α.

Remark 3.3 It should be noted that if one directly uses the conventional Holm procedure with the critical constants $\alpha_i = \alpha/(2n-i+1), i = 1, \ldots, 2n$ for testing the $2n$ hypotheses, then the critical constants corresponding to the first n most significant hypotheses will be always less than or equal to α/n. However, for the above modified Holm procedure, the critical constants corresponding to the first n most significant hypotheses are generally much larger than α/n. The main reason why such procedure works well is that the $2n$ tested hypotheses have some structural relationship: they can be arranged as n pairs of one true and one false null hypotheses. For each pair of hypotheses, the sum of their corresponding p-values is equal to one. Thus, for each pair of hypotheses, when one hypothesis is significant, another one is impossible to be significant. The introduced modified Holm procedure has fully exploited the above facts and hence is more powerful than the conventional Holm procedure.

Remark 3.4 It should be noted that when testing n null hypotheses, the critical constants of the above modified Holm procedure are slightly less than those of the usual Holm procedure, thus this procedure can also strongly control $\text{FWER}_{\mathcal{F}_2}$ at level α. Therefore, if we use separate analysis approach to test $\mathcal{F}$ by applying separately this procedure to test $\mathcal{F}_1$ and $\mathcal{F}_2$ at level $\alpha/2$, then the $\text{FWER}_{\mathcal{F}}$ is strongly controlled at level α.

4 Controlling the mdFWER under dependence

In this section, we will discuss how to control the $\text{FWER}_{\mathcal{F}_1}$ under three different types of dependence: within- and between-block dependence, and positive dependence.

4.1 Controlling the $\text{FWER}_{\mathcal{F}_1}$ under block dependence

Suppose that $\mathcal{F}_1 = \{H_1, \ldots, H_{2n}\}$ can be organized as b subfamilies $\mathcal{F}_{1i}, i = 1, \ldots, b$, each of which have n_i pairs of null hypotheses, (H_j, H_{n+j}), with $\sum_{i=1}^{b} n_i = n$. If the test statistics corresponding to the true null hypotheses within each subfamily are mutually independent and no dependence assumption is made on the test statistics corresponding to the true null hypotheses between these subfamilies, then we term such dependence structure as *between-block dependence*. In contrast, if the test statistics corresponding to the true null hypotheses between the subfamilies are mutually independent and no dependence assumption is made on the test statistics corresponding to the true null hypotheses within each subfamily, then we term such dependence structure as *within-block dependence*.

By using the stepdown procedure introduced in Theorem 3.2, a method for testing $\mathcal{F}_1$ can be constructed as follows:

- For $i = 1, \ldots, b$, use a stepdown procedure in Theorem 3.2 for testing $\mathcal{F}_{1i}$ at level $\beta_i = n_i\alpha/n$.
- Let K_i be the corresponding set of rejected null hypotheses for testing $\mathcal{F}_{1i}$. Reject all null hypotheses in $\bigcup_{i=1}^{b} K_i$.

Under the assumption of between-block dependence, through Theorem 3.2, the FWER of the stepdown procedure for testing $\mathcal{F}_{1i}$, $\text{FWER}_{\mathcal{F}_{1i}}$, satisfies $\text{FWER}_{\mathcal{F}_{1i}} \leq n_i\alpha/n$. Thus, the overall FWER of the above method for testing $\mathcal{F}_1$ satisfies

$$\text{FWER}_{\mathcal{F}_1} \leq \sum_{i=1}^{b} \text{FWER}_{\mathcal{F}_{1i}} \leq \sum_{i=1}^{b} \frac{n_i\alpha}{n} = \alpha.$$

Therefore, we have the following result:

Theorem 4.1 *Consider the multiple testing method defined as above. Under assumption A.1-A.2 and the assumption of between-block dependence, this method strongly controls the* $FWER_{\mathcal{F}_1}$ *at level* α.

Remark 4.1 When the number of subfamilies b is equal to n, that is, each subfamily has only one pair of hypotheses, the above method reduces to a modified Bonferroni procedure with the critical constant $\alpha/(n+\alpha)$, which strongly controls the $\text{FWER}_{\mathcal{F}_1}$ under arbitrary dependence. When there is only one subfamily, the above method reduces to the stepdown procedure in Theorem 3.2, which strongly controls the $\text{FWER}_{\mathcal{F}_1}$ under independence. Finally, we should point out that the critical constants $\frac{n_i\alpha/n}{n_i-j+1+n_i\alpha/n}$ of the stepdown procedure used in the above method are almost always larger than or equal to α/n, which implies that the method is generally more powerful than the usual Bonferroni procedure with the critical constant α/n.

Remark 4.2 When the test statistics corresponding to the above b subfamilies are within-block dependent rather than between-block dependent, we can reorganize these b subfamilies as $n_{\max}$ new subfamilies such that the corresponding test statistics are between-block dependent, where $n_{\max} = \max\{n_i : i = 1, \ldots, b\}$. Then, we can apply the above suggested method to test $\mathcal{F}_1$ based on these reorganized subfamilies and it results in the corresponding $\text{FWER}_{\mathcal{F}_1}$ is controlled at level α.

4.2 Controlling the $\text{FWER}_{\mathcal{F}}$ under positive dependence

In this subsection, we discuss how to control the $\text{FWER}_{\mathcal{F}}$ rather than $\text{FWER}_{\mathcal{F}_1}$ under positive dependence. First, reorganize $\mathcal{F}$ as two new subfamilies, $\mathcal{F}_1' = \{H_{i1} : i = 1, \ldots, n\}$ and $\mathcal{F}_2' = \{H_{ij} : i = 1, \ldots, n, j = 2, 3\}$. Thus, for each $i = 1, 2$, the test statistics corresponding to the null hypotheses in $\mathcal{F}_i'$ are positively dependent (which is not the case for $\mathcal{F}_1$, leading to the the current division into subfamilies).

Based on the conventional Hochberg procedure (Hochberg, 1988), which is the stepup procedure with critical constants $\alpha_i = \alpha/(n - i + 1), i = 1, \ldots, n$ that strongly controls the FWER at level α under positive dependence, a method for simultaneously testing $\mathcal{F}$ can be constructed as follows:

- Use the Hochberg procedure to test $\mathcal{F}_1'$ at level $\alpha/2$.
- Use the Hochberg-type procedure with the critical constants $\alpha_i = \frac{\alpha}{n-\lfloor (i+1)/2 \rfloor+1}, i = 1, \ldots 2n$, to test $\mathcal{F}_2'$ at level $\alpha/2$.
- For $i = 1, 2$, let K_i be the corresponding set of rejected null hypotheses for testing

$\mathcal{F}_i'$. Reject all null hypotheses in $K_1 \bigcup K_2$.

Note that for H_{i2} and H_{i3}, their corresponding p-values are the same. Thus, when we apply the aforementioned Hochberg-type procedure to test $\mathcal{F}_2'$ at level $\alpha/2$, it is equivalent to apply the conventional Hochberg procedure with the critical constants $\alpha_i = \alpha/(n-i+1), i = 1, \dots, n$ to test H_{i2}'s or H_{i3}'s. Then, the corresponding $\text{FWER}_{\mathcal{F}_2'}$ is controlled at level $\alpha/2$. (Of course, α could be split into β and $\alpha - \beta$, but for simplicity $\beta = \alpha/2$.) Hence,

$$\text{FWER}_{\mathcal{F}} \le \text{FWER}_{\mathcal{F}_1'} + \text{FWER}_{\mathcal{F}_2'} \le \alpha/2 + \alpha/2 = \alpha.$$

Theorem 4.2 *Consider the multiple testing method defined as above. Under assumption A.1 and the assumption of positive dependence in the sense of (4), this method strongly controls the $FWER_{\mathcal{F}}$ at level α.*

5 Controlling the mixed directional FDR under independence and dependence

In this section, we discuss how to control the $\text{FDR}_{\mathcal{F}_1}$ under the same settings as in the last two sections.

5.1 On the $\text{FDR}_{\mathcal{F}_1}$ control under independence

Consider the BH-type procedure with the critical constants $\alpha_i = i\alpha/n, i = 1, \dots, n$ for testing $\mathcal{F}_1 = \{H_1, \dots, H_{2n}\}$. Note that $P_{n+i} = 1 - P_i$ for each $i = 1, \dots, n$; thus, among the $2n$ corresponding p-values, there are n p-values larger than or equal to 0.5. Therefore, for the BH-type procedure, it is sufficient to only define its first n critical constants while testing those $2n$ null hypotheses. Under assumptions A.1 and A.3, for any given parameter vector $\theta = (\theta_1, \ldots, \theta_n)$, we have

$$
\begin{aligned}
\mathrm{FDR}_{\mathcal{F}_1}(\theta) &= E\left\{\frac{V_1}{R_1 \vee 1}\right\} \\
&= \sum_{i=1}^{n} E_\theta \left\{\frac{I\{H_i \text{ rejected}\}}{R_1 \vee 1}\right\} \\
&= \sum_{i=1}^{n}\sum_{r=1}^{n} \frac{1}{r} P_\theta\{R_1 = r,\ H_i \text{ rejected}\} \\
&= \sum_{i=1}^{n}\sum_{r=1}^{n} \frac{1}{r} \Pr_\theta \left(R_1 = r, P_i \le \frac{r}{n}\alpha\right) \\
&= \sum_{i=1}^{n}\sum_{r=1}^{n} \frac{1}{r} \Pr_\theta \left(R_1^{\{-i,-(n+i)\}} = r-1, P_i \le \frac{r}{n}\alpha\right) \\
&\le \sum_{i=1}^{n}\sum_{r=1}^{n} \frac{\alpha}{n} \Pr_\theta \left(R_1^{\{-i,-(n+i)\}} = r-1\right) \\
&= \alpha. \qquad (16)
\end{aligned}
$$

Here, $R_1^{\{-i,-(n+i)\}}$ is the number of rejected null hypotheses by using the stepup procedure with the critical values $j\alpha/n, j = 2, \ldots, n$ to simultaneously test the $2(n-1)$ null hypotheses $H_1, \ldots, H_{2n}$ excluding the pair of null hypotheses (H_i, H_{n+i}). In (16), the inequality follows from assumptions A.1 and A.3 and the fact that $P_{n+i} = 1 - P_i$. Therefore, the following conclusion holds.

Theorem 5.1 *Consider the stepup procedure with the critical constants* $\alpha_i = \frac{i\alpha}{n}, i = 1, \ldots, n$. *Under assumptions A.1 and A.3, the procedure strongly controls the* $\mathrm{FDR}_{\mathcal{F}_1}$ *at level* α.

Remark 5.1 Note that assumption A.2 is not used. In fact, the result holds without the parametric model assumptions used in much of this paper. Indeed, all that is assumed is the availability of p-values P_i for testing some parameter $\theta_i = 0$ and their independence. Of course, we must have $P_{n+i} = 1 - P_i$, but this is a natural requirement when constructing two one-sided p-values.

Remark 5.2 When $\theta = 0$, the inequality in (16) becomes an equality. Thus the above

BH-type procedure cannot be improved in terms of its BH-type critical constants while maintaining the control of the $\text{FDR}_{\mathcal{F}_1}$.

5.2 On the $\text{FDR}_{\mathcal{F}_1}$ control under between-block dependence

Suppose that $\mathcal{F}_1 = \{H_1, \ldots, H_{2n}\}$ can be organized as b subfamilies $\mathcal{F}_{1i}, i = 1, \ldots, b$, each of which have n_i pairs of null hypotheses (H_j, H_{n+j}) with $\sum_{i=1}^{b} n_i = n$. Assume that the test statistics corresponding to those subfamilies satisfy the condition of between-block dependence.

By using the BH-type procedure in Theorem 5.1, a method for simultaneously testing $\mathcal{F}_1$ can be constructed as follows:

- For each given $i = 1, \ldots, b$, use the BH-type procedure to test $\mathcal{F}_{1i}$ at level $n_i\alpha/n$.
- Let K_i be the corresponding set of rejected null hypotheses for testing $\mathcal{F}_{1i}$. Reject all null hypotheses in $\bigcup_{i=1}^{b} K_i$.

Under the assumption of between-block dependence, through Theorem 6, the FDR of the BH-type procedure for testing subfamily $\mathcal{F}_{1i}$ at level $n_i\alpha/n$ satisfies $\text{FDR}_{\mathcal{F}_{1i}} \leq n_i\alpha/n$. Thus, the overall FDR of the above method for $\mathcal{F}_1$ satisfies

$$\text{FDR}_{\mathcal{F}_1} \leq \sum_{i=1}^{b} \text{FDR}_{\mathcal{F}_{1i}} \leq \sum_{i=1}^{b} \frac{n_i\alpha}{n} = \alpha.$$

Theorem 5.2 *Consider the multiple testing method defined as above. Under assumption A.1 and the assumption of between-block dependence, this method strongly controls the $FDR_{\mathcal{F}_1}$ at level α.*

Theorem 5.2 implies Theorem 5.1. When $b = 1$, it reduces to Theorem 5.1.

5.3 On the $\text{FDR}_{\mathcal{F}_1}$ control under within-block dependence

Suppose that $\mathcal{F}_1 = \{H_1, \ldots, H_{2n}\}$ can be organized as b subfamilies $\mathcal{F}_{1i}, i = 1, \ldots, b$, each of which have n_i pairs of null hypotheses (H_j, H_{n+j}) with $\sum_{i=1}^{b} n_i = n$. Assume that the test statistics corresponding to those subfamilies satisfy the condition of within-block dependence. For $i = 1, \ldots, b$, let $\widetilde{P}_i$ denote the smallest one among the n_i pairs of p-values corresponding to the n_i pairs of null hypotheses in $\mathcal{F}_{1i}$. Note that there are exactly n true null hypotheses in $\mathcal{F}_1$, by exploiting the information in a two-stage BH-type procedure introduced in Guo and Sarkar (2014), a method for simultaneously testing $\mathcal{F}_1$ is constructed as follows:

- Order these smallest p-values $\widetilde{P}_i, i = 1, \ldots, b$ as $\widetilde{P}_{(1)} \leq \cdots \leq \widetilde{P}_{(b)}$, and find $B = \max\{1 \leq i \leq b : \tilde{P}_{(i)} \leq i\alpha/n\}$.
- In each subfamily $\mathcal{F}_{1i}$, reject those null hypotheses whose corresponding p-values are less than or equal to $B\alpha/n$.

By using the same arguments as in Guo and Sarkar (2012), we can show that the above method strongly controls the $\text{FDR}_{\mathcal{F}_1}$ at level α. Therefore, we have the following result.

Theorem 5.3 *Consider the multiple testing method defined as above. Under assumption A.1 and the assumption of within-block dependence, this method strongly controls the $FDR_{\mathcal{F}_1}$ at level α.*

Theorem 5.3 implies Theorem 5.1. When $b = n$, it reduces to Theorem 5.1.

5.4 On the $\text{FDR}_{\mathcal{F}_1}$ control under positive dependence

Suppose that the test statistics $T_i, i = 1, \ldots, n$ are positively dependent in the sense of (4). Then, for each $j = 1, 2$, the test statistics corresponding to the true null hypotheses $H_{ij}, i = 1, \ldots, n$ are also positively dependent. For $j = 1, 2$, let $\mathcal{F}_{1j} = \{H_{ij}, i = 1, \ldots, n\}$ and n_{1j} denote the number of true nulls in $\mathcal{F}_{1j}$. Note that there are exactly n true null hypotheses in $\mathcal{F}_1 = \mathcal{F}_{11} \bigcup \mathcal{F}_{12}$, thus $n_{11} + n_{12} = n$. By using the similar idea due to Benjamini and Yekutieli (2005), a method for testing $\mathcal{F}_1$ can be constructed as follows:

- For $j = 1, 2$, use the BH procedure with the critical constants $\alpha_i = i\alpha/n, i = 1, \ldots, n$

to test $\mathcal{F}_{1j}$ at level α.

- Let K_j be the corresponding set of rejected null hypotheses for testing $\mathcal{F}_{1j}$. Reject all null hypotheses in $K_1 \bigcup K_2$.

By using the result in Benjamini and Yekutieli (2001) and Sarkar (2002) on the FDR control of the BH procedure under positive dependence, we have

$$\mathrm{FDR}_{\mathcal{F}_1} \leq \mathrm{FDR}_{\mathcal{F}_{11}} + \mathrm{FDR}_{\mathcal{F}_{12}} \leq \frac{n_{11}\alpha}{n} + \frac{n_{12}\alpha}{n} = \alpha.$$

The equality follows from the fact that $n_{11} + n_{12} = n$.

Theorem 5.4 *Consider the multiple testing method defined as above. Under assumption A.1 and the assumption of positive dependence in the sense of (4), this method strongly controls the FDR$_{\mathcal{F}_1}$ at level α.*

6 Concluding remarks

In this paper, several approaches, methods, and results are presented addressing the multiple testing problem of accounting for both type 1 and type 3 errors. Many of the results required the assumption of independence, which is quite strong, though we have weakened this assumption as well. The problem of directional error control has proven to be quite challenging, and though we do not consider the dependent case more fully, it is hoped to consider this important problem in future work.

7 Acknowledgments

The research of the first author is supported by NSF Grant DMS-1006021 and DMS-1309162. The research of the second author is supported by NSF Grant DMS-0707085.

8 Appendix

8.1 Proof of Lemma 1

Since the family of densities $f_{i,\theta}(\cdot)$ satisfies the assumption of MLR, we have that, for any given $\theta_1 > \theta_0$ and $x_1 > x_0$,

$$\frac{f_{i,\theta_1}(x_1)}{f_{i,\theta_0}(x_1)} \geq \frac{f_{i,\theta_1}(x_0)}{f_{i,\theta_0}(x_0)}. \tag{17}$$

By multiplying both sides of (17) by $f_{i,\theta_0}(x_0)$ and then integrating over x_0 from $-\infty$ to x_1, one obtains

$$\frac{f_{i,\theta_1}(x)}{f_{i,\theta_0}(x)} \geq \frac{F_{i,\theta_1}(x)}{F_{i,\theta_0}(x)} . \tag{18}$$

Similarly, one obtains

$$\frac{1-F_{i,\theta_1}(x)}{1-F_{i,\theta_0}(x)} \geq \frac{f_{i,\theta_1}(x)}{f_{i,\theta_0}(x)}. \tag{19}$$

Consider the functions $G_1(x) = \frac{F_{i,\theta_1}(x)}{F_{i,\theta_0}(x)}$ and $G_2(x) = \frac{1-F_{i,\theta_1}(x)}{1-F_{i,\theta_0}(x)}$. It is easy to check by using (18) and (19) that $G_1'(x) \geq 0$ and $G_2'(x) \geq 0$. Then, $G_1(x)$ and $G_2(x)$ are both non-decreasing in x. First, assume $\theta_i > 0$, so that $P_i = F_{i,0}(T_i)$. Thus, for any $0 \leq p < p' \leq 1$,

$$\Pr_{\theta_i}\left\{P_i \leq p \middle| P_i \leq p'\right\} = \frac{F_{i,\theta_i}(x)}{F_{i,\theta_i}(x')} \leq \frac{F_{i,0}(x)}{F_{i,0}(x')} = \frac{p}{p'} = \Pr_{\theta_i=0}\left\{P_i \leq p \middle| P_i \leq p'\right\}, \tag{20}$$

where $x = F_0^{-1}(p)$ and $x' = F_0^{-1}(p')$. In (20), the inequality follows from the fact that $G_1(x)$ is non-decreasing in x and the second equality follows from assumption A.1. By using similar arguments, we can prove that (20) also holds when $\theta_i < 0$. Hence, the desired result follows. ■

References

[1] Bauer, P., Hackl, P., Hommel, G. and Sonnemann, E. (1986). Multiple testing of pairs of one-sided hypotheses. *Metrika* **33** 121-127.

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

[3] Benjamini, Y. and Braun, H. (2002). John W. Tukey's contributions to multiple comparisons. *Ann. Statist.* **30** 1576–1594.

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

[5] Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J Amer. Statist. Assoc.* **100** 71-93.

[6] Finner, H. (1994). Testing multiple hypotheses: general theory, specific problems, and relationships to other multiple decision procedures. Habilitationsschrift, Fachbereich IV Mathematik, Univ. Trier.

[7] Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *Ann. Statist.* **27** 274-289.

[8] Finner, H. and Gontscharuk, V. (2009). Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *J. Roy. Statist. Soc., Ser. B* **71** 1031–1048.

[9] Guo, W. (2009). A note on adaptive Bonferroni and Holm procedures under dependence. *Biometrika*, **96**, 1012-1018.

[10] Guo, W. and Sarkar, S. (2014). Adaptive controls of FWER and FDR under block dependence. Working Paper.

[11] Guo, W., Sarkar, S. and Peddada, S. (2010). Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics* **66** 485-492.

[12] Hochberg Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800-802.

[13] Holm, S. (1979a). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65-70.

[14] Holm, S. (1979b). *A stagewise directional test based on t statistics.* Statistical Research Report No.1979-3, Institute of Mathematics, Chalmers University of Technology, Gothenburg, Sweden.

[15] Holm, S. (1981). A stagewise directional test for the normal regression situation. In B. Bereanu, S. Grigorescu, M. Josifescu, & T. Postelnicu (Eds.), *Proceedings of the Sixth Conference on Probability Theory*, 103–106, Bucharest, Romania.

[16] Jones, L. V. and Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods* **5** 411–414.

[17] Liu, W. (1997). Control of directional errors with step-up multiple tests. *Statist. Probab. Lett.* **31** 239-242.

[18] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30** 239-257.

[19] Sarkar, S. K., Sen, P. K. and Finner, H. (2004). On two results in multiple testing. In *Recent Developments in Multiple Comparisons.* IMS Lectures Notes-Monograph Series, 47, Y. Benjamini, F. Bretz and S. Sarkar, eds. 89-99, Institute of Mathematical Statistics, Beachwood.

[20] Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *Ann. Statist.* **8** 1342-1347.

[21] Shaffer, J. P. (2002). Multiplicity, directional (type III) errors, and the null hypothesis. *Psychological Methods* **7** 356-369.

[22] Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science* **6** 100–116.

[23] Westfall, P. H., Bretz, F. and Tobias, R. D. (2013). Directional error rates of closed testing procedures. *Statistics in Biopharmaceutical Research* **5** 345–355.

Department of Mathematical Sciences

New Jersey Institute of Technology

E-mail: (wenge.guo@njit.edu)

Departments of Statistics and Economics

Stanford University

E-mail: (romano@stanford.edu)