
Paper summary

FDR control in ordered hypothesis testing

Jinxi Liu

Contents

1	Lynch, Gavin, et al. "The control of the false discovery rate in fixed sequence multiple testing." arXiv preprint arXiv:1611.03146 (2016).	2
1.1	A conventional fixed sequence method that stops testing once an acceptance occurs	2
1.1.1	Under arbitrary dependence	2
1.1.2	Under negative dependence	2
1.2	Fixed Sequence Procedures that Allow More Acceptances	2
1.2.1	Under arbitrary dependence	3
1.2.2	Under the independence assumption	3
1.3	Data Driven Ordering	3
1.4	Assumption	4
1.4.1	Distribution of P-values	4
1.4.2	Positive regression dependence(PRDS)	4
1.4.3	Negative Association	4
2	G'Sell, Max Grazier, et al. "Sequential selection procedures and false discovery rate control." Journal of the royal statistical society: series B (statistical methodology) 78.2 (2016): 423-444.	5
2.1	ForwardStop Method	5
2.2	StrongStop	6
2.3	Discussion	7
3	Lynch, Gavin, and Wenge Guo. "On Procedures Controlling the FDR for Testing Hierarchically Ordered Hypotheses." arXiv preprint arXiv:1612.04467 (2016).	8
3.1	Notation and definition	8
3.2	Assumptions	9
3.3	A generalized stepwise procedure which generalizes the usual stepwise procedure to the case where each hypothesis is tested with a different set of critical constants.	9
3.4	Hierarchical testing procedures which control the FDR under various forms of dependence such as positive dependence and block dependence.	10
3.4.1	General Hierarchical Testing Procedure	10
3.4.2	Procedure under Positive Dependence	11
3.5	Discussion	11

1 Lynch, Gavin, et al. "The control of the false discovery rate in fixed sequence multiple testing." arXiv preprint arXiv:1611.03146 (2016).

In this paper, the authors focus on the fixed sequence structure where the testing order of the hypotheses has been strictly specified in advance. They develop step-down FDR controlling methods that fully exploit the fixed sequence structural information, in which hypotheses are tested from lowest-ranked to highest-ranked. They first consider a conventional fixed sequence method that stops testing once an acceptance occurs, and develop such a method controlling the FDR under both arbitrary and negative dependencies. To account for any potential mistakes in the ordering of the tests, they extend the conventional fixed sequence method to one that allows more but a given number of acceptances.

1.1 A conventional fixed sequence method that stops testing once an acceptance occurs

The authors present the developments of two simple conventional fixed sequence procedures controlling the FDR under both arbitrary dependence and negative dependence conditions on the p-values.

1.1.1 Under arbitrary dependence

Consider a conventional fixed sequence procedure with critical constants

$$\alpha_i^{(1)} = \min\left(\frac{m\alpha}{m-i+1}, 1\right), i = 1, \dots, m.$$

- (i) This procedure strongly controls the FDR at level α under arbitrary dependence of the p-values.
- (ii) One cannot increase even one of the critical constants $\alpha_i^{(1)}$, $i = 1, \dots, m$, while keeping the remaining fixed without losing control of the FDR. This is true even when \vec{P} is assumed to be PRDS on \vec{P}_0 .

1.1.2 Under negative dependence

When the p-values are negatively associated, the critical constants of the conventional fixed sequence procedure can be further improved as in the following: The conventional fixed sequence method with critical constants

$$\alpha_i^{(2)} = \frac{i\alpha}{1 + (i-1)\alpha}, i = 1, \dots, m.$$

strongly controls the FDR at level α when the p-values are negatively associated on the true null p-values.

1.2 Fixed Sequence Procedures that Allow More Acceptances

There is a potential for loss of power in a conventional fixed sequence multiple testing method if the ordering of the hypotheses, particularly for the earlier ones, does not match with that of their true effect sizes, potentially leading to some earlier hypothesis being accepted and the follow-up hypotheses having no chance to be tested. To mitigate that, we consider generalizing the conventional fixed sequence multiple testing to one that allows more than one but a pre-specified number of acceptances, and develop such generalized fixed sequence multiple testing methods controlling the FDR under both arbitrary dependence and independence.

Fixed sequence method stopping on the k_{th} acceptance

- 1 If $P_1 < \alpha_1$, then reject H_1 ; otherwise, accept H_1 . If $k > 1$ or H_1 is rejected, then continue to test H_2 ; otherwise stop.
- 2 If $P_i < \alpha_i$, then reject H_i ; otherwise, accept H_i . If the number of accepted hypotheses so far is less than k , then continue to test H_{i+1} ; otherwise stop.

1.2.1 Under arbitrary dependence

The fixed sequence method stopping on the k_{th} acceptance with critical constants

$$\alpha_i^{(3)} = \begin{cases} \frac{\alpha}{k}, & \text{if } i = 1, \dots, k \\ \frac{(m-k+1)\alpha}{(m-i+1)k}, & \text{if } i = k+1, \dots, m \end{cases}$$

strongly controls the FDR at level α arbitrary dependence of p-values.

1.2.2 Under the independence assumption

Under certain distributional assumptions on the p-values, the critical constants of the above procedure can potentially be improved. In particular, we have the following result.

Consider a fixed sequence method stopping on the k_{th} acceptance with critical constants

$$\alpha_i^{(4)} = \frac{(r_{i-1} + 1)\alpha}{k + (i - k)\alpha}, i = 1, \dots, m,$$

where r_i is the number of rejections among the first i tested hypotheses, with $r_0 = 0$, for $i = 1, \dots, m$. This procedure strongly controls the FDR at level α if the true null p-values are mutually independent and are independent of the false null p-values.

1.3 Data Driven Ordering

The applicability of the aforementioned fixed sequence methods depends on availability of natural ordering structure among the hypotheses. When the hypotheses cannot be pre-ordered, one can use pilot data available to establish a good ordering among the hypotheses in some cases. For example, in replicated studies, the hypotheses for the follow-up study can be ordered using the data from the primary study. However, when prior information is unavailable ordering information can usually be assessed from the data itself.

Assume that the variables of interest are $X_i, i = 1, \dots, m$, with n independent observations X_{i1}, \dots, X_{in} on each X_i . An ordering statistic, Y_i , and a test statistic, T_i , are determined for each $i = 1, \dots, m$. The Y_i 's are used to order all of the hypotheses H_1, \dots, H_m , T_i is used to test the hypothesis $H_i; i = 1, \dots, m$, and P_i is the corresponding p-value. In addition, Y_i is chosen such that it is independent of the T_i 's under H_i and tends to be larger as the effect size increases.

Example 2: Two sample T-test. Consider testing $H_i : \mu_i^{(1)} = \mu_i^{(2)}$ against $H_i' : \mu_i^{(1)} \neq \mu_i^{(2)}$ simultaneously using $n = n_1 + n_2$ data vectors. Suppose $X_{ij}^{(l)}, j = 1, \dots, n_l$, follows a $N(\mu_i^{(l)}, \sigma^2)$ distribution, for $l = 1, 2$. Then, the hypotheses can be tested using the two-sample t-test statistics T_i and are ordered through the values of the total sum of squares, which is $Y_i = \sum_{l=1}^2 \sum_{j=1}^{n_l} X_{ij}^{(l)2} / n$, for $i = 1, \dots, m$. The rationale behind this is independence between the Y_i 's and T_i under H_i , and the following result: $E[Y_i] = (n-1)\sigma^2 + n_1 n_2 (\mu_i^{(1)} - \mu_i^{(2)})^2 / n$.

The fixed sequence multiple testing procedures introduced above can still strongly control the FDR at level α .

1.4 Assumption

1.4.1 Distribution of P-values

They assume that the true null p -values, \hat{P}_i , for $i = 1, \dots, m_0$, are marginally distributed as follows:

$$Pr\left(\hat{P}_i \leq p\right) \leq p \text{ for any } p \in (0, 1).$$

1.4.2 Positive regression dependence (PRDS)

The vector of p -values \vec{P} is PRDS on the vector of null p -values $\vec{P}_0 = (\hat{P}_1, \dots, \hat{P}_{m_0})$ if for every increasing set D and for each $i = 1, \dots, m_0$, the conditional probability $Pr\left(\vec{P} \in D | \hat{P}_i = p\right)$ non-decreasing in p .

1.4.3 Negative Association

The vector of p -values \vec{P} is negatively associated with null p -values if for each $i = 1, \dots, m_0$, the following inequality holds:

$$\begin{aligned} &Pr\left(\hat{P}_i \leq p_{\mu_i}, P_j \leq p_j, j = 1, \dots, m, \text{ with } j \neq \mu_i\right) \leq \\ &Pr\left(\hat{P}_i \leq p_{\mu_i}\right) Pr\left(P_j \leq p_j, j = 1, \dots, m, \text{ with } j \neq \mu_i\right) \end{aligned}$$

for all fixed p_j 's.

Several multivariate distributions possess the conventional negative association property, including multivariate normal with non-positive correlation, multinomial, dirichlet, and multivariate hypergeometric. Independence is a special case of negative dependence.

2 G'Sell, Max Grazier, et al. "Sequential selection procedures and false discovery rate control." *Journal of the royal statistical society: series B (statistical methodology)* 78.2 (2016): 423-444.

The problem of FDR control for ordered hypotheses arises naturally when implementing variable selection using a path-based a path-based regression algorithm.; examples of such algorithms include forward stepwise regression and least angle regression. In this paper, the authors consider a multiple hypothesis testing setting where the hypotheses are ordered and one is only permitted to reject an initial contiguous block, H_1, \dots, H_k , of hypotheses. A rejection rule in this setting amounts to a procedure for choosing the stopping point k . In the ordered setting, a valid rejection rule is a function of p_1, \dots, p_m that returns a cutoff \hat{k} such that hypotheses $H_1, \dots, H_{\hat{k}}$ are rejected. The False Discovery Rate (FDR) is defined as $E[V(\hat{k})/\max(1, \hat{k})]$, where $V(\hat{k})$ is the number of null hypotheses among the rejected hypotheses $H_1, \dots, H_{\hat{k}}$.

The key difference between the setup of Benjamini and Hochberg (1995) and their problem is that, in the former, the rejection set R can be arbitrary, whereas here they must always reject the first k hypotheses for some k . For example, even if the p -value corresponding to the third hypothesis is very small, they cannot reject the third hypothesis unless they also reject the first and second hypotheses.

Their proposed methods start by transforming the sequence of p -values p_1, \dots, p_m into a monotone increasing sequence of statistics $0 \leq q_1 \leq \dots, q_m \leq 1$. They then prove that they achieve ordered FDR control by applying the original Benjamini-Hochberg procedure on the monotone test statistics q_i .

2.1 ForwardStop Method

Suppose that we could transform our p -values p_1, \dots, p_m into statistics $q_1 < \dots < q_m$, such that the q_i behaved like a sorted list of p -values. Then, we could apply the BH procedure on the q_i , and get a rejection set R of the form $R = 1, \dots, k$.

Under the global null where $p_1, \dots, p_m \stackrel{iid}{\sim} U([0, 1])$, we can achieve such a transformation using the Rényi representation theorem (Rényi, 1953). Rényi showed that if Y_1, \dots, Y_m are independent standard exponential random variables, then

$$\left(\frac{Y_1}{m}, \frac{Y_1}{m} + \frac{Y_2}{m-1}, \dots, \sum_{i=1}^m \frac{Y_i}{m-i+1} \right) \stackrel{d}{=} E_{1,m}, E_{2,m}, \dots, E_{m,m},$$

where the $E_{i,m}$ are exponential order statistics, meaning that the $E_{i,m}$ have the same distribution as a sorted list of independent standard exponential random variables.

In our context, let

$$\begin{aligned} Y_i &= -\log(1 - p_i), \\ z_i &= \sum_{j=1}^i Y_j / (m - j + 1), \text{ and} \\ q_i &= 1 - e^{-Z_i}. \end{aligned}$$

Under the global null, the Y_i are distributed as independent exponential random variables. Thus, by Rényi representation, the Z_i are distributed as exponential order statistics, and so the q_i are distributed like uniform order statistics.

This argument suggests that in an ordered selection setup, we should reject the first \hat{k}_F^q hypotheses where

$$\hat{k}_F^q = \text{MAX} \left\{ k : q_k \leq \frac{\alpha k}{m} \right\}$$

Procedure 1 (ForwardStop). *Let $p_1, \dots, p_m \in [0, 1]$, and let $0 < \alpha < 1$. We reject hypotheses $1, \dots, \hat{k}_F$, where*

$$\hat{k}_F = \max \left\{ k \in \{1, \dots, m\} : \frac{1}{k} \sum_{i=1}^k Y_i \leq \alpha \right\},$$

and

$$Y_i = -\log(1 - p_i).$$

Suppose that we have p -values $p_1, \dots, p_m \in (0, 1)$, a subset $N \subseteq \{1, \dots, m\}$ and are null i.e., independently drawn from $U([0, 1])$. Then, the ForwardStop procedure controls FDR at level α , meaning that

$$E \left[\left| \{1, \dots, \hat{k}_F\} \cap N \right| / \max \{ \hat{k}_F, 1 \} \right] \leq \alpha$$

Forwardstop first transforms the p -values, and then sets the rejection threshold at the largest k for which the first k transformed p -values have a small enough average. If the first p -values are very small, then ForwardStop will always reject the first hypotheses regardless of the last p -values. As a result, the rule is moderately robust to potential misspecification of the null distribution of the p -values at high indexes.

2.2 StrongStop

In the previous procedure, the ordered test statistics Z_i was created by by summing transformed p -values starting from the first p -value. This choice was in some sense arbitrary. Under the global null, we could just as well obtain uniform order statistics q_i by summing from the back:

$$\begin{aligned} \tilde{Y}_i &= -\log(1 - p_i), \\ \tilde{z}_i &= \sum_{j=i}^m Y_j / j, \text{ and} \\ \tilde{q}_i &= 1 - e^{-\tilde{z}_i}. \end{aligned}$$

Procedure 2 (StrongStop). Let $p_1, \dots, p_m \in [0, 1]$, and let $0 < \alpha < 1$. We reject hypotheses $1, \dots, \hat{k}_S$, where

$$\hat{k}_S = \max \left\{ k \in \{1, \dots, m\} : \tilde{q}_k \leq \frac{\alpha k}{m} \right\},$$

and \tilde{q}_k is as defined above.

Unlike ForwardStop, this new procedure needs to look at the p -values corresponding to the last hypotheses before it can choose to make any rejections. This can be a liability if we do not trust the very last p -values much. Looking at the last p -values can however be useful if the model is correctly specied, as it enables us to strengthen our control guarantees: StrongStop not only controls the FDR, but also controls the FWER. **Note this guarantees only hold when the non-null p -values all precede the null ones.**

Suppose that we have p -values $p_1, \dots, p_m \in (0, 1)$, the last $m - s$ of which are null (i.e., independently drawn from $U([0, 1])$). Then the above rule controls the FWER at level α , meaning that

$$P \left[\hat{k}_S > s \right] \leq \alpha.$$

StrongStop comes with a stronger guarantee than ForwardStop. Provided that the non-null p-values precede the null ones, it not only controls the FDR, but also controls the Family-Wise Error Rate (FWER) at level α . The main weakness of StrongStop is that the decision to reject at k depends on all the p-values after k . If the very last p-values are slightly larger than they should be under the uniform hypothesis, then the rule suffers a considerable loss of power.

2.3 Discussion

A major advantage of both ForwardStop and StrongStop is that these procedures seek the largest k at which an inequality holds, even if the inequality may not hold for some index l with $l < k$. This property enables them to get past some isolated large p -values for the early hypotheses, thus resulting in a substantial increase in power. The methods in this article require that the null p-values be independent.

3 Lynch, Gavin, and Wenge Guo. "On Procedures Controlling the FDR for Testing Hierarchically Ordered Hypotheses." arXiv preprint arXiv:1612.04467 (2016).

In many problems involving the testing of multiple hypotheses, the hypotheses have an intrinsic, hierarchical structure such as a tree-like or graphical structure. In general, hierarchical testing typically occurs while testing hierarchically structured hypotheses where, upon the rejection of one hypothesis, followup hypotheses are to be tested. For instance, Heller et al. (2009) introduced a hierarchical testing approach for analyzing microarray data where individual genes were grouped into gene sets. The gene sets were tested and upon successfully rejecting a gene set, the associated individual genes were tested.

In this paper, the authors propose new FDR controlling methods for testing hierarchically ordered hypotheses under various dependencies.

To their knowledge, the procedures are the first procedures developed for testing hierarchically ordered hypotheses with proven control of the FDR under dependence structures other than independence.

3.1 Notation and definition

Suppose there are m hypotheses H_1, \dots, H_m to be tested that are organized hierarchically in a tree-like structure where each hypothesis can have several child hypotheses but at most one parent hypothesis.

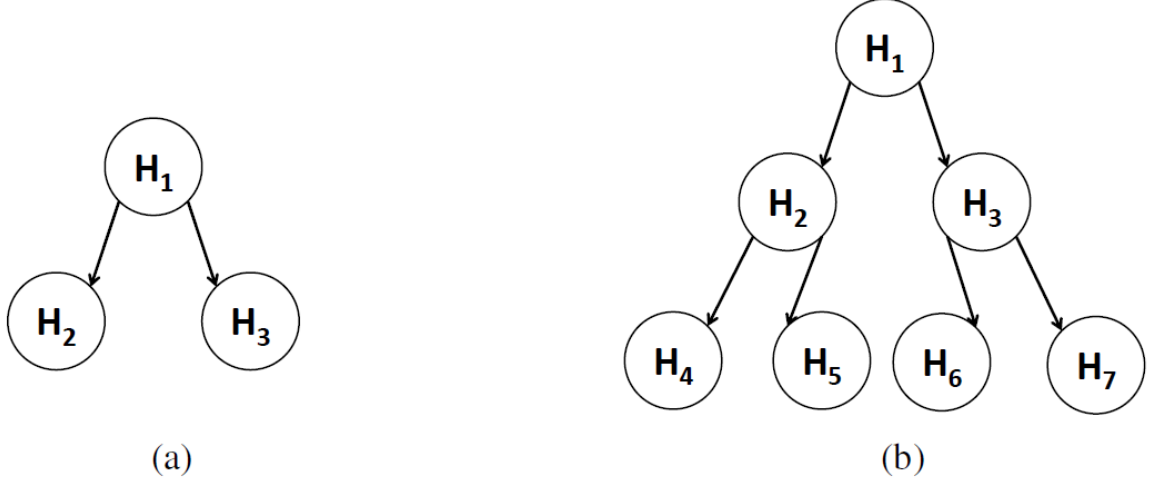


Figure 1: (a) An example of a hierarchical structure with 3 hypotheses for which H_2 and H_3 are only tested if H_1 is rejected. (b) An example of a hierarchical structure with 7 hypotheses for which H_2 and H_3 are only tested if H_1 is rejected, H_4 and H_5 are only tested if H_2 is rejected, and H_6 and H_7 are only tested if H_3 is rejected.

1. Let $M = \{H_1, \dots, H_m\}$ be the set of the m tested hypotheses.
2. Let $T: \{0, \dots, m\} \rightarrow \{0, \dots, m\}$ be a function that takes an index of a hypothesis and returns the

index of the parent hypothesis with $T(0) = 0$.

3. D_i is the set of all ancestor hypotheses of H_i , which includes H_i .
4. d_i is the cardinality of D_i . The depth of H_i in the hierarchy is defined as d_i .
5. Let D be the maximum depth of the m hypotheses to be tested.
6. M_i is the set of all descendant hypotheses of H_i , which also includes H_i .
7. Let m_i be the cardinality of M_i .
8. If $m_i = 1$, then H_i has no children and it is referred to as a leaf hypothesis.
9. Let l be the number of leaf hypotheses in the whole hierarchy and l_i be the number of leaf hypotheses in the subtree under H_i .
10. Hypotheses are grouped into D families by depth where family d contains all hypotheses with depth d . That is, $F_d = \{H_i \in M, : d_i = d\}$.

For example, in Figure 1 (a), $T(2) = T(3) = 1$ and H_2 and H_3 are leaf hypotheses. In Figure 1(b), $T(6) = T(7) = 3$; $D_6 = \{H_1, H_3, H_6\}$, $M_2 = \{H_2, H_4, H_5\}$, and $F_3 = \{H_4, H_5, H_6, H_7\}$.

3.2 Assumptions

Throughout this paper the authors make use of the following basic assumption regarding marginal p -values: for any p -value P_i with H_i being true,

$$Pr(P_i \leq p) \leq p \text{ for any } 0 \leq p \leq 1.$$

They consider several types of joint dependence throughout this paper: arbitrary dependence, positive dependence, and block dependence. Under arbitrary dependence, the p -values are not known to have any specific type of dependence structure. Positive dependence and block dependence are characterized by the following assumptions.

Positive Dependence Assumption *For any coordinatewise non-decreasing function of the p -values ψ , $E(\psi(P_1, \dots, P_m) | P_i \leq p)$ is non-decreasing in p for each p -value P_i such that H_i is true.*

Block Dependence Assumption *For each $d = 1, \dots, D$, the p -values corresponding to the hypotheses in F_d are independent of the p -values corresponding to the hypotheses not in F_d .*

Positive Dependence Assumption is slightly more relaxed than the condition of positive regression dependence on a subset (PRDS) introduced in Benjamini and Yekutieli (2001). Block Dependence Assumption only characterizes the joint dependence of the p -values across families but does not describe the joint dependence within families.

3.3 A generalized stepwise procedure which generalizes the usual stepwise procedure to the case where each hypothesis is tested with a different set of critical constants.

Most existing multiple testing procedures are stepwise methods which are based on the ordered p -values. Typically the rejection thresholds of a stepwise procedure are based on a sequence of non-decreasing critical constants but in this paper, for convenience, the authors instead test the hypotheses using a non-decreasing, non-negative function $\alpha_0 : \{0, \dots, m + 1\} \rightarrow R$ called a critical function where $\alpha_0(0) = 0$. For example, the critical function of the BH procedure is $\alpha_0(r) = r\alpha/m$.

Given m non-decreasing critical functions $\alpha_i(r), i = 1, \dots, m$, the proposed generalized stepwise procedure rejects H_i if $P_i \leq \alpha_i(R)$ for each $i = 1, \dots, m$ where R is determined as follows.

For the generalized stepup procedure,

$$R = \max \left\{ 0 \leq r \leq m : r \leq \sum_{i=1}^m I\{P_i \leq \alpha_i(r)\} \right\},$$

For the generalized stepdown procedure,

$$R = \min \left\{ 1 \leq r \leq m + 1 : r > \sum_{i=1}^m I\{P_i \leq \alpha_i(r)\} \right\} - 1,$$

Example 1 Consider a weighted multiple testing problem where H_i has corresponding weight $w_i, i = 1, \dots, m$. A weighted stepwise procedure with the critical function $\alpha_0(r)$ tests H_i based on weight-adjusted p -values P_i/w_i instead of P_i . This is equivalent to a generalized stepwise procedure with the critical functions $\alpha_i(r) = w_i \alpha_0(r), i = 1, \dots, m$ so that the weighted stepwise procedure can be regarded as a special case of the generalized stepwise procedure.

Example 2 Fixed sequence procedures assume the testing order of the hypotheses has been specified a-priori and that H_i is not tested unless H_1, \dots, H_{i-1} have all been rejected. Lynch et al. (2016) showed that the fixed sequence procedure that rejects H_i when $P_i \leq m\alpha/(m - r + 1)$ controls the FDR at level α under arbitrary dependence. This procedure is a special case of the generalized stepdown procedure with critical functions $\alpha_i(r) = I\{r \geq i\}m\alpha/(m - r + 1)$.

The authors also present an efficient algorithm for finding the number of rejections by the generalized stepwise procedure. The algorithm is particularly useful when the number of hypotheses is very large.

3.4 Hierarchical testing procedures which control the FDR under various forms of dependence such as positive dependence and block dependence.

3.4.1 General Hierarchical Testing Procedure

The tested hypotheses are arranged into D families, F_1, \dots, F_D , where F_d is the family of hypotheses with depth d . Given m non-decreasing critical functions $\alpha_i(r); i = 1, \dots, m$, the hypotheses are tested as follows.

General Hierarchical Testing Procedure

1. Test F_1 by using the generalized stepup procedure with critical functions $\alpha_i(r), H_i \in F_1$. Let S_1 be the set of rejected hypotheses and $R(F_1)$ be the number of rejected hypotheses in F_1 . Test F_2 .
2. To test F_d , use the generalized stepup procedure with critical functions

$$\alpha_i^*(r) = I\{H_{T(i)} \text{ is rejected}\} \alpha_i \left(r + \sum_{j=1}^{d-1} R(F_j) \right), H_i \in F_d.$$

Let S_d be the set of rejected hypotheses and $R(F_d)$ be the number of rejected hypotheses in F_d . Test F_{d+1} .

3. The set of rejected hypotheses is $\cup_{d=1}^D S_d$ and the total number of rejections is $R = \sum_{d=1}^D R(F_d)$.

The above procedure is termed as a hierarchical testing procedure since the procedure will accept any hypothesis whose parent hypothesis has been accepted.

3.4.2 Procedure under Positive Dependence

Under positive dependence assumption, the hierarchical testing procedure with critical functions

$$\alpha_i(r) = \frac{l_i \alpha}{l} \frac{m_i + r - 1}{m_i}, i = 1, \dots, m,$$

strongly controls the FDR at level α . Consider the special case when there is no hierarchical

3.5 Discussion

An interesting finding of this research is that when the hierarchy takes on some special configurations, our procedures reduce to the existing FDR controlling procedures. For example, when there is no hierarchical structure, our proposed procedures reduce to the BH procedure and the Benjamini-Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001). When the hierarchy takes on a fixed sequence structure, our procedures are equivalent to the fixed sequence procedures in Lynch et al. (2016). This shows that our procedures are the combination of stepwise and fixed sequence methods.