New Step Down Procedures for Control of the Familywise Error Rate

---

A Dissertation
Submitted to
the Temple University Graduate Board

---

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Zijiang Yang
January, 2009

ABSTRACT

New Step Down Procedures for Control of the Familywise Error Rate

Zijiang Yang

DOCTOR OF PHILOSOPHY

Temple University, January, 2009

Professor Sanat K. Sarkar, Chair

The main research topic in this dissertation is the development of the closure method of multiple testing procedures. Considering a general procedure that allows the underlying test statistics as well as the associated parameters to be dependent, we first propose a step-down procedure controlling the FWER, which is defined as the probability of committing at least one false discovery.

Holm (1979) first proposed a step-down procedure for multiple hypothesis testing with a control of the familywise error rate (FWER) under any kind of dependence. Under the normal distributional setup, Seneta and Chen (2005) sharpened the Holm procedure by taking into account the correlations between the test statistics. In this dissertation, the Seneta-Chen procedure is further modified yielding a more powerful FWER controlling procedure. We then advance our research and propose another step-down procedure to control the generalized FWER (k-FWER), which is defined as the probability of making at least k false discoveries. We compare our proposed k-FWER procedure with the Lehmann and Romano (2005) procedure. The proposed k-FWER procedure is more powerful, particularly when there is a strong dependence in

the tests.

When the proportion of true null hypotheses is expected to be small, the traditional tests are usually conservative by a factor associated with π_0, which is the proportion of true null hypotheses among all null hypotheses. Under independence, two procedures controlling the FWER and the k-FWER are proposed in this dissertation. Simulations are carried out to show that our procedures often provide much better FWER or k-FWER control and power than the traditional procedures.

Key words: Multiple Comparisons; Familywise Error Rate; Generalized Familywise Error Rate; Closure Method; Step-down Test

ACKNOWLEDGEMENTS

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Sanat Sarkar, who has supported me throughout my thesis with his patience and knowledge whilst allowing me the room to work in my own way. I attribute my degree to his encouragement and effort and without him this thesis, too, would not have been completed or written. One simply could not wish for a better or friendlier supervisor.

I would also like to thank my committee members, Dr. Damaraju Raghavarao, Dr. Francis Hsuan and Dr. Steven Chang for their constructive advice in all the stages of my research work. Their efforts were indispensable for the completion of my thesis research.

I also express my gratitude to Ms. Jingjing Chen at Clinforce and Dr. Bingming Yi at GlaxoSmithKline for allowing me to have a flexible work schedule. Without their kind support, my research work can not go smoothly.

The Department of Statistics has provided the support I have needed to produce and complete my thesis and this research is supported by the NSF Grants DMS-0306366 and DMS-0603868.

Finally, I thank my wife for supporting me throughout all my studies at Temple University and for providing a home in which to complete my writing up.

To my parents and my wife,

Fajin Yang, Xiaoying Tong and Yanping Liu

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In today's environment, multiple testing techniques are widely accepted and used in a variety of industries, such as finance, accounting and biology. Due to the advances in computing facilities and data collection methods, statistical analysts are being confronted with huge data sets. For example, a fund manager's portfolio construction requires him to choose several companies out of thousands. As another example, in gene microarray studies, thousands of genes or proteins are being monitored simultaneously with the support of modern experimental technology, such as DNA microarray and protein array.

Comparative studies are very commonly employed in these fields. For example, while there are several different techniques used to build a fund portfolio, historical stock returns are often the most powerful evidence that identifies the companies that significantly outperform the market or a benchmark. Similarly, among several different statistical objectives of microarray experiments, [see Amaratunga and Cabrera (2004) for more details], one specific objective is to compare expression levels of a set of genes across two or more conditions; in particular, to identify genes that are significantly differentially expressed

across these conditions.

To pinpoint a specific application of multiple testing in finance, let's consider evaluating a large number of industrial companies, say, 400 companies in S&P500 with a goal to find out which companies, if there are any, outperformed the market in a certain period. One possible approach to make the multiple comparisons is to address each comparison separately by a suitable procedure. That is, in order to detect differences between 400 stocks' mean returns and the mean return of the market index, one could perform 400 separate t-tests, each at certain level α. Such multiple t-tests are in fact used quite frequently in practice. This kind of approach, however, does not account for the *multiplicity effect*. Therefore, one cannot and should not draw a conclusion like: Company A, B and C simultaneously outperformed the market (and start to build a portfolio consists of A, B and C), even if their p-values are all less than α. It is clear that even none of the companies really outperform the market, about 5 percent, or 20 companies will have exceptional performance by chance alone. Thus, when we consider to construct a portfolio, we must take into account of those 20 "lucky" companies, and should not include them in our portfolio. In the words of Grinold and Kahn (2000) "The fundamental goal of performance analysis is to separate skill from luck. But, how do you tell them apart? None of the successful managers will admit to being lucky; all of the unsuccessful managers will cite bad luck."

As an example in microarray analysis, experiments are often conducted to compare the expression levels of several genes in cancerous liver cells versus those in normal liver cells, and vice versa. Again, multiplicity problems should be taken care of. The task of detecting outstanding companies or differentially

expressed genes is performed by using multiple hypothesis testing, with the null hypothesis corresponding to each company not being able to outperform the benchmark, or each gene representing no change in expression levels for that gene.

1.1 Some Basics in Multiple Hypothesis Testing

Suppose that we have n null hypotheses $H_1, \ldots, H_n$ to be tested simultaneously. For instance, in a microarray experiment, with n being the number of genes and H_i as the null hypothesis representing no difference in expression levels for the ith gene, we consider developing a multiple testing procedure to identify the genes that are differentially expressed. Or, in a portfolio selection, with n being the number of companies and H_i as the null hypothesis representing no exceptional performance for the ith company, we consider developing a multiple testing procedure to identify the outperforming companies.

In multiple hypothesis testing, the main problem is to determine a rule to specify what decision should be made for each null hypothesis, based on test statistics or p-values associated with the null hypotheses. The rule is determined based on the idea of using a procedure that leads us to the right decisions with high probability and a control of a suitably defined error rate. In the case of a single hypothesis, this problem is quite simple. A good procedure would be the one that controls Type-I error rate, which is the probability of making false rejections, while minimizing Type-II error rate (or maximizing the power), which is the probability of making false acceptances (or making correct

Table 1.1: Outcomes of simultaneously testing n hypotheses

	Accepted	Rejected	Total
True Null	U	V	n_0
False Null	T	S	n_1
Total	A	R	n

rejections). This task is much more complicated when dealing with multiple hypotheses. First, as we mentioned before, the simple approach of addressing each comparison separately would lead to too many false rejections, unless the test statistics of all hypotheses are perfectly correlated. Second, there is no unique extension of the concept of Type-I error rate from single testing to multiple testing. Finally, one has to take into consideration of the correlations between the tests if they exist.

Table 1.1 summarizes all possible outcomes in a multiple testing procedure. Different types of error rate could then be defined based on this table.

Note that the total number of hypothesis n is fixed and known; the number of true and false null hypotheses n_0 and n_1 are also fixed but always unknown; Random variables A and R are the total number of acceptances and the total number of rejections respectively, and these two variables are observable; V is the number of false discoveries (Type-I errors) and T is the number of false non-discoveries (Type-II errors); Similarly, U is the number of correct acceptances and S is the number of correct rejections. These four variables are not observable.

1.1.1 Type-I Error Rates in Multiple Hypothesis Testing

There are several different measures of error rates in multiple testing. The following are traditional error rates, [see, e.g., Hochberg and Tamhane (1987)]:

- Per-Family Error Rate: The expected number of false rejections, i.e.

$$\text{PFER} = E(V).$$

- Per-Comparison Error Rate: The expected proportion of false rejections, i.e.

$$\text{PCER} = E(V)/n.$$

- Familywise Error Rate: The probability of having at least one false rejection, i.e.

$$\text{FWER} = \Pr(V \geq 1).$$

The FWER has been the most commonly used among the above error rates. However, with large number of hypotheses, as encountered in many modern statistical investigations, procedures controlling it become too stringent, resulting in conservative procedures with inadequate power. To overcome this problem, the following alternative measures have been introduced.

- False Discovery Rate (FDR): The expected proportion of false discoveries among all rejections. Let

$$\text{FDP} = \begin{cases} \frac{V}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

be the False Discovery Proportion. Then, the FDR is defined as

$$\text{FDR} = E(\text{FDP}) = E\left(\frac{V}{R} \mid R > 0\right) \cdot \Pr(R > 0). \tag{1.1}$$

- Positive False Discovery Rate (pFDR): The expected proportion of false discoveries among all rejections given there is at least one rejection, i.e.

$$\text{pFDR} = E(\text{FDP} \mid R > 0) = E\left(\frac{V}{R} \mid R > 0\right). \tag{1.2}$$

- Generalized Familywise Error Rate (k-FWER): The probability of having at least k false rejections for a pre-specified integer k, i.e.

$$k\text{-FWER} = \Pr(V \geq k).$$

  Note that when $k = 1$, the k-FWER reduces to the original FWER.

- The Exceedance Probability of False Discovery Proportion (γ-ExFDP): The probability of FDP exceeding $\gamma \in (0, 1)$, i.e.

$$\text{ExFDP}(\gamma) = \Pr(\text{FDP} > \gamma).$$

- Generalized False Discovery Rate (k-FDR): The expected proportion of k or more false discoveries among all rejections, where k is pre-specified, i.e.

$$k\text{-FDR} = E(k\text{-FDP}), \tag{1.3}$$

  where

$$k\text{-FDP} = \begin{cases} \frac{V}{R} & \text{if } V \geq k; \\ 0 & \text{if } V < k. \end{cases}$$

  Note that when $k = 1$, it reduces to the original FDR.

The concept of FDR has been introduced by Benjamini and Hochberg (1995), which is more powerful than the concept of the FWER. However, as Storey (2002) argues, there are some difficulties with this notion of FDR. To cope with these difficulties, he [Storey (2002) and Storey (2003)] introduced the notion of positive FDR (pFDR) by deleting the second term in the FDR definition (1.1). Of course, the pFDR cannot always be controlled, as it is equal to 1 when all hypotheses are true. Storey (2002) proposed an estimation based approach to controlling it through a suitable estimate of it for a fixed rejection region. Similar idea can be used to control the FDR also.

In many applications, particularly where n is large, one might be willing to tolerate more than one false rejection and seeks to control at least k, rather than at least one, false rejections. This will improve the power of a procedure to detect more false null hypotheses. The concepts of k-FWER, γ-ExFDP and k-FDR have been introduced as appropriate measures of error rates in these situations. Korn et al. (2004) first considered using the k-FWER and the γ-ExFDP. Sarkar (2007) proposed the idea of controlling the k-FDR. It is a less conservative notion than the k-FWER and is a natural generalization of the idea of improving the FWER using the FDR.

It is important to note that both the FWER and FDR and their generalizations are defined under a fixed configuration, although unknown, of true and false null hypotheses. A *Strong Control* is achieved for a rate when it is controlled at a pre-specified level under any configuration of true and false null hypotheses; whereas, a *Weak Control* is achieved when it is controlled when all the null hypotheses are assumed to be true.

In general, controlling a rate in weak sense is unsatisfactory because it is unrealistic for the overall null hypothesis to be true. The configuration of the true and false null hypotheses is usually unknown. Therefore, strong control is always desired as it ensures the control under any kind of configuration. It is, however, important to note that often procedures with a strong control of the FWER are constructed based on those controlling the FWER in a weak sense using the closure principle of Marcus et al. (1976) [to be discussed in Section 2.1.5]. In this proposal, by saying that a rate is controlled we mean that is strongly, unless noted otherwise.

From the definitions given above, it is easy to see that:

$$\begin{aligned} &\text{PCER} \leq \text{FDR} \leq \text{FWER} \leq \text{PFER} \\ &k\text{-FDR} \leq k\text{-FWER} \leq \text{FWER} \\ &k\text{-FDR} \leq \text{FDR} \\ &\text{FDR} \leq \text{pFDR} \end{aligned} \tag{1.4}$$

The equation (1.4) presents relative conservativeness of the different error rates discussed above. For example, if a procedure controls the FWER, it will also control the FDR and the k-FWER for $k > 1$. Since the FDR when all the null hypotheses are true is same as the FWER in a weak sense, a procedure with a control of the FDR controls the FWER weakly. Moreover, an FDR procedure also controls the k-FDR. The smaller error rates provide less conservative procedures in the sense of allowing more rejections.

1.1.2 Type-II Error Rates and Power

Having had different multiple testing procedures controlling an overall measure of Type-I errors, it is natural to compare them in terms of a performance

measuring criterion. In single testing testing, this criterion is usually the power or the Type-II error rate. In multiple testing, however, the concept of *power* is not uniquely defined. The following are measures of power proposed in the literature.

- Probability of rejecting all false null hypothesis:

  $$\Pr(S = n_1) = \Pr(T = 0).$$

  This power might be too small to compare, especially when the number of hypotheses is large.

- Probability of rejecting at least one false null hypothesis:

  $$\Pr(S > 0) = \Pr(T \neq n_1).$$

  This power might be too large to compare, especially when the number of hypothesis is large.

- Average Power, the expected proportion of correct rejections among all false null hypotheses (AvePower):

  $$\text{AvePower} = \begin{cases} \frac{E(S)}{n_1} & \text{if } n_1 > 0 \\ 1 & \text{if } n_1 = 0. \end{cases} \tag{1.5}$$

- Correct Non-discovery Rate (CNR):

  $$\text{CNR} = E\left(\frac{U}{A} \mid A > 0\right) \cdot \Pr(A > 0) + \Pr(A = 0).$$

- Number of rejections: This measure is simple, and is often used when comparing two procedures controlling the same error rate. The procedure with larger total number of rejections is said to be more powerful than the other.

The CNR is related to the concept of False Non-discovery Rate (FNR) as follows:

$$\begin{aligned} \text{FNR} &= 1 - \text{CNR} \\ &= E\left(\tfrac{T}{A} \mid A > 0\right) \cdot \Pr(A > 0), \end{aligned} \tag{1.6}$$

which is an analogue of the FDR in terms of the Type-II errors and was defined by Genovese and Wasserman (2002), and independently by Sarkar (2004) (who calls it the False Negative Rate).

1.2 Types of Multiple Testing Procedure

There are different types of multiple testing procedure. Let $P_1, \ldots, P_n$ be the p-values corresponding to the null hypotheses $H_1, \ldots, H_n$, respectively. Let these p-values be sorted as $P_{1:n} \leq \cdots \leq P_{n:n}$, with the corresponding hypotheses being $H_{1:n}, H_{2:n}, \ldots, H_{n:n}$, respectively. Given a set of critical values $\alpha_1 \leq \cdots \leq \alpha_n$, the following are the different types of *Stepwise Procedure*.

- Step-Down Procedures (SDP): Find $i_0 = \min\{1 \leq i \leq n : P_{i:n} \geq \alpha_i\}$. Reject the hypotheses: $H_{1:n}, \ldots, H_{i_0-1:n}$. If the minimum does not exist, reject all the null hypotheses.

- Step-Up Procedures (SUP): Find $i_0 = \max\{1 \leq i \leq n : P_{i:n} \leq \alpha_i\}$. Reject the hypotheses: $H_{1:n}, \ldots, H_{i_0:n}$. If the maximum does not exist, accept all the null hypotheses.

- Generalized Step-Up-Down Procedures (SUDP): An SUDP of order r starts from $P_{r:n}$. If $P_{r:n} > \alpha_r$, the procedure accepts $H_{r:n}, \ldots, H_{n:n}$ and continues testing the remaining ones in a Step-Up manner using the

corresponding p- and critical values. If $P_{r:n} \leq \alpha_r$, the procedure rejects $H_{1:n}, \ldots, H_{r:n}$ and continues testing the remaining ones in a Step-Down manner using the corresponding p- and critical values.

- Single-Step Procedures: A stepwise procedure reduces to a single-step procedure when the critical values are all same.

When $r = 1$ (or n), the SUDP of order r reduces to the ordinary Step-Down (or Step-Up) procedure. The SDP was first introduced by Miller (1966) and is widely used in multiple hypothesis testing. The SUDP was introduced by Tamhane et al. (1998); also see Sarkar (2002b).

A *Two-Stage Procedures* procedure is a kind of stepwise procedure, but the critical values are not pre-determined. The critical values in a stage are defined based on the number of rejections or acceptances in the previous stage.

Two-stage procedures have been used in the context of FDR. More specifically, it was used to estimate n_0, the number of true null hypotheses, in the first stage and then use it to modify the critical values of an FDR procedure in the second stage [Storey (2002), Storey et al. (2004) and Sarkar (2008b)]

Generally, stepwise and two-stage procedures are more powerful than the corresponding single-step or single-stage procedure. Procedures like these controlling different error rates will be discussed in details in Chapter 2.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we will briefly review the procedures which control the FWER and the k-FWER. By controlling the FWER or the k-FWER , we generally mean in the strong sense, unless it is specified otherwise. We will first state some general theories and definitions useful in multiple comparison procedures, some of which are outlined in Hochberg and Tamhane (1987). Then we will review the current available procedures for controlling the FWER and the k-FWER.

2.1 Some General Theories

2.1.1 Coherence and Consonance

Both coherence and consonance were introduced by Gabriel (1969). Any MTP for a hierarchical family of hypotheses $\{H_i, i \in I\}$ is generally required to possess the following logical consistency property: For any pair of hypotheses (H_i, H_j) such that H_j implies H_i, if H_j is not rejected then H_i is also not

rejected. That is, if H_j is not rejected, any of its components should not be rejected. This requirement is called *coherence*. An MTP that satisfies this requirement is called *coherent*. A coherent MTP avoids the inconsistency of rejecting a hypothesis without also rejecting all hypotheses implying it.

For a hierarchical family of hypotheses $\{H_i, i \in I\}$, *consonance* refers to the property that whenever any nonminimal H_i is rejected, at least one of its components is also rejected. An MTP that has this property is called *consonant*.

The lack of consonance does not imply logical contradictions as the lack of coherence does. This is because the failure to reject a hypothesis is not usually interpreted as its acceptance. Also sometimes the failure to reject a component of H_i may be due to noninclusion of enough components of H_i in the family. Therefore, while coherence is an essential requirement for multiple testing, consonance is only a desirable property. For example, we may reject $\theta_1 = \theta_3$, but fail to reject both $\theta_1 = \theta_2$ and $\theta_2 = \theta_3$.

2.1.2 Union-Intersection Method

Roy (1953) proposed a heuristic method of constructing a test of any hypothesis H_0 that can be expressed as an intersection of a family of hypothesis, which is referred to as Union-Intersection (UI) Method.

Suppose that $H_0 = \bigcap_{i \in I} H_i$ where I is an arbitrary index set. Further suppose that a suitable test of each H_i is available. Then, according to the UI method, the rejection region for H_0 is given by the union of rejection regions for the $H_i, i \in I$, that is, H_0 is rejected if and only if at least one H_i is rejected.

If the UI test on H_0 is of level α, then all inferences derived from it have

the FWER strongly controlled at α. An MTP derived in this manner from a UI test is referred to as a *UI procedure*.

A general presentation of the UI method for constructing a single-step test procedure can be given as follows: Let I_{min} be the index set of the minimal hypotheses in $\{H_i, i \in I\}$ and let $I_{min}^{(j)}$ be the index set of the minimal components of a nonminimal H_j. We assume that every nonminimal H_j could be expressed as:

$$H_j = \bigcap_{i \in I_{min}^{(j)}} H_i, \qquad j \in I - I_{min}$$

Let φ_i be the indicator test function of H_i when data is observed, that is, $\varphi_i = 1$ if H_i is rejected and 0 if not. Given the indicator test functions φ_i for all $i \in I_{min}$, the UI test for any nonminimal H_j is given by

$$\varphi_j = \max_{i \in\ I_{min}^{(j)}} \{\varphi_i\} = 1 - \prod_{i \in\ I_{min}^{(j)}} \{1 - \varphi_i\} \tag{2.1}$$

A UI procedure is derived by first constructing tests for H_i, $i \in I_{min}$, and then obtaining tests of nonminimal H_j by using (2.1). If the critical region of each H_i for $i \in I_{min}$ is of the form $Z_i > \xi$, then from (2.1) we see that the test statistic for any nonminimal H_j will be $Z_j = \max_{i \in I_{min}}^{(j)} Z_i$ and its critical region would be $Z_i > \xi$. The Z_j's obtained in this manner are referred to as UI statistics. They are monotone because if H_j implies H_i then $I_{min}^{(i)} \subseteq I_{min}^{(j)}$ and therefore $Z_j = \max_{l \in I_{min}^{(j)}} Z_l \geq Z_i = \max_{l \in I_{min}^{(i)}} Z_l$. This implies coherence.

UI procedure also validates the approach of sequential p-values. Assume that the marginal distributions and the joint distributions of the test statistics are all the same under null hypothesis, i.e., multivariate normal distribution with a common correlation, it makes sense that larger p-values would be rejected if and only if we have rejected all the smaller ones.

2.1.3 Free Combinations Condition

Holm (1979) proposed the following property, which is referred to as the *Free Combinations Condition*: Suppose that for any $\{P \in \{1, 2, \ldots, n\}\}$ the set of parameter points for which all H_i's, $i \in P$, is true and all H_j's, $j \notin P$, are false is non-empty for any choice of P. This means that every partition of the n hypotheses into two subsets such that the hypotheses in one subset are true and those in the other subset are false is possible for at least some point in the parameter space. To be short, the condition means that it is not possible that the truth of two hypotheses could imply the truth or falseness of a third hypothesis. That is, any subset of the null hypotheses could be the set of true hypotheses. Please note that the family of all pairwise comparisons does not satisfy this condition. For example, $\theta_1 = \theta_2$ and $\theta_2 = \theta_3$ would imply $\theta_1 = \theta_3$. Therefore, $\theta_1 = \theta_2$, $\theta_2 = \theta_3$ and $\theta_1 \neq \theta_3$ is empty. The *free combinations condition* is a conservative condition, therefore, when it does not hold, we may be able to improve our test procedures. In this article, we assume this condition holds, unless noted otherwise.

2.1.4 Bonferroni and Kounias Inequalities

Kounias (1968) derived the following inequality to obtain an upper bound of the probability of a union:

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \min\left\{\sum_{i=1}^{n} P_i - \max_{k=1,2,\ldots,n} \sum_{i=1, i\neq k}^{n} P_{k,i}, 1\right\} \tag{2.2}$$

where A_i represents a random event and $P_i = \Pr(A_i)$ and $P_{k,i} = \Pr(A_k \cap A_i)$.

Kounias Inequality can be viewed as an adjustment of the traditional Bonferroni Inequality by implementing second-degree associations. In particular,

when A_i denotes $P_i < \alpha_0$, (2.2) becomes:

$$\begin{aligned}
\Pr\left(\bigcup_{i=1}^{n} P_i < \alpha_0\right) = \; & \Pr(P_{(1)} < \alpha_0) \\
\leq \; & \min\left\{ \sum_{i=1}^{n} \Pr(P_i < \alpha_0) \quad - \right. \\
& \left. \max_{k=1,2,\ldots,n} \sum_{i=1, i\neq k}^{n} \Pr(P_k < \alpha_0, P_i < \alpha_0), 1 \right\} \\
\leq \; & \sum_{i=1}^{n} \Pr(P_i < \alpha_0) \qquad \text{Bonferroni Inequality}
\end{aligned}$$

Therefore, Kounias Inequality clearly attained a better upper bound than Bonferroni Inequality.

2.1.5 Closure Method

Marcus et al. (1976) proposed a general method for constructing step-down test procedures. This method is referred to as *closure method* and the corresponding procedures are referred to as *closed testing procedures*.

Let $H_1, H_2, \ldots, H_n$ be a finite family of hypotheses. Form the closure of this family by taking all non-empty intersections $H_I = \cap_{i\in I} H_i$ for $I \subseteq \{1, 2, \ldots, n\}$. If a level-$\alpha$ test is available for each H_I, then the closed testing procedure rejects any hypothesis H_I if and only if every H_J is rejected by its corresponding level-α test for all $J \supseteq I$. Marcus et al. (1976) proved that the closed testing procedure strongly controls the FWER at level α.

Now consider a closed testing procedure that uses UI statistics for testing all intersection hypotheses H_I, as defined above. Since the UI tests have the property that whenever any intersection hypothesis H_I is rejected, at least one H_i implied by H_I is rejected. Therefore, in order to make a decision on H_i, one does not need to test on all the intersections containing H_i; one just need to test the most insignificant intersection, since if it is rejected, all the other

intersections will be rejected. Further, if all the level-α tests are a single-step test, we will have a shortcut version of the closure method, which is the so-called step-down procedure. Holm (1979) proposed this shortcut version of the closed testing procedure. See more details in Section 2.3.2.

2.1.6 PRDS Property

An n-dimensional random vector $X = (X_1, \ldots, X_n)$ or the corresponding multivariate distribution is said to be *positive regression dependent on a subset* (PRDS) $M : \{X_i, i \in M\}$, where $M \subseteq \{1, \ldots, n\}$, if $\Pr(\mathbf{X} \in \mathbf{C}|X_i)$ is non-decreasing (non-increasing) in X_i for each $i \in M$, for any increasing (decreasing) set $\mathbf{C}$. A set $\mathbf{C}$ is increasing (decreasing) if and only if $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{C}$ implies that $\mathbf{x}' = (x_1', \ldots, x_n') \in \mathbf{C}$ for any $x_i \leq x_i'$ $(x_i \geq x_i')$, $i = 1, \ldots, n$.

The PRDS property is a relaxed form of the positive regression dependency property. The latter means that for any increasing set $\mathbf{C}$, $\Pr(\mathbf{X} \in \mathbf{C}|X_i = x_i, i = 1, \ldots, n)$ is nondecreasing in $(x_1, \ldots, x_n)$ as stated in Sarkar (1969). In PRDS the conditioning is on one variable only each time and required to hold only for a subset of the variables. The multivariate normal distribution with nonnegative correlations is PRDS on any subset. Actually, multivariate distribution of independent test statistics and multivariate F are all PRDS on any subset also. See Benjamini and Yekutieli (2001) and Sarkar (2002b) for more details.

A lot of FWER and FDR controlling procedures require the test statistics to have a multivariate distribution that is PRDS on $I_0 = \{1, \ldots, n_0\}$, that is, on the subset of statistics $(X_1, \ldots, X_{n_0})$ corresponding to the true null

hypotheses. See more details in Section 2.2.

2.2 Multiple Testing Procedures Controlling the FWER Weakly

As mentioned in Section 1.1.1, the weak FWER control is achieved when it is controlled when all the null hypotheses are assumed to be true. That is, given the null hypotheses: $H_1, \ldots, H_n$ with the corresponding p-values, $P_1, \ldots, P_n$, the FWER is weakly controlled if:

$$\Pr(\text{rejecting at least one } H_i \mid H_0 : \bigcap_{i=1}^{n} H_i) \leq \alpha.$$

While there are several such procedures, we will describe a few in the following:

2.2.1 LSD Procedure

One of the earliest MTP is Fisher's LSD test or *protected least significance difference test* as referred to in the literature. This procedure is for making inferences on pairwise comparisons.

The LSD Procedure is a two-step procedure. In the first step, an α-level F-test is applied to test the overall H_0: $\bigcap_{i=1}^{n} H_i$. If it is not significant, then the procedure stops without making any further inferences on individual pairwise comparisons. Otherwise, each pairwise comparisons is tested by an α-level t-test.

For LSD procedure, the preliminary F-test on H_0 provides a weak control for the FWER.

2.2.2 Simes' Procedure

Simes (1986) originally proposed this procedure to control the Type-I error of the global hypothesis.

Simes' procedure rejects $H_0 : \bigcap_{i=1}^{n} H_i$ if $P_{(i)} < i\alpha/n$ for any $i = 1, \ldots, n$.

Simes proved that this procedure controls the Type-I error rate under the independence of the p-values. Since the Type-I error rate is the FWER given all hypotheses are true, Simes' procedure weakly controls the FWER. Based on a simulation study, he also conjectured that when the p-values are correlated, the procedure controls the Type-I error rate under a large variety of distributions. Sarkar and Chang (1997) and Sarkar (1998) proved the Simes' conjecture for random variables with common marginal and MTP_2 property.

When the overall hypothesis H_0 is rejected, Simes also suggested a rule to make decisions over individual hypothesis: Reject $H_{(1)}, \ldots, H_{(j)}$, where $j = \max\{k : P_{(k)} \leq k\alpha/n\}$. This is a step-up procedure based on the critical values: $\alpha_i = i\alpha/n$ for $i = 1, \ldots, n$. However, this step-up procedure can only weakly control the FWER and cannot strongly control FWER even with independent test statistics [see Hommel (1988)]. However, as Simes' procedure provides a level-α test for every intersection of the family, it leads to a procedure with a strong control of the FWER using the closure method. This is what Hommel (1988) did in his procedure. Details of the Hommel's procedure are given in Section 2.3.3.

2.2.3 Cai and Sarkar's Procedures

Cai and Sarkar (2008) and Cai and Sarkar (2006) modified Simes' procedure weakly controlling the FWER under independence and positive dependence,

respectively.

Under independence, Cai and Sarkar (2008) defined their modified Simes' critical values as follows:

Let β_0 be an arbitrary number between 0 and α/n inclusive. Then the following α_i's will control the FWER exactly at α when the test statistics are iid under null hypotheses:

$$\alpha_i = (n-i+1)\beta_i, \quad \text{where } \beta_i = \frac{\beta_0}{2} + \sqrt{\frac{\beta_0^2}{4} + \frac{(n-i)(\alpha - n\beta_0)}{n(n-1)(n-i+1)}}$$

By choosing $\beta_0 = \alpha/n$, the procedure becomes the original Simes' procedure. By choosing $\beta_0 = 0$, we will have:

$$\alpha_i = \sqrt{\frac{(n-i)(n-i+1)\alpha}{n(n-1)}}$$

for $i = 1, \ldots, n$. Cai and Sarkar's critical values are larger than Simes' except for the smallest one. Simes (1986) proved that under independence assumption, his procedure controls the FWER at exact α for the overall null hypothesis. Therefore, the procedure can not be uniformly improved under independence. Cai and Sarkar's procedure with independence assumption sacrifices the smallest critical value for the compensation of substantial increase in the other ones. Thus, the procedure is still a level-α test, but is to be found more powerful in most cases.

In Cai and Sarkar (2006) Simes' procedure is improved under positive dependence. As Simes' procedure is proved to be conservative in this case (see Sarkar (1998)), Cai and Sarkar's procedure is uniformly more powerful than the original Simes. A disadvantage of their procedure is that it requires the same technique of calculating critical values as in Dunnett and Tamhane (1992), see

Section 2.3.6 for more details, thus it works well only for small n. Moreover, no theoretical proof exists for the monotonicity of these critical values.

Kwong et al. (2002) also proposed a procedure to improve Simes test under positive dependence by only increasing the largest critical value.

2.3 Multiple Testing Procedures Controlling the FWER Strongly

Assume that $\hat{P}_1, \ldots, \hat{P}_{n_0}$ are the p-values corresponding to the true null hypotheses. Then, the FWER of a step-down procedure satisfies the following:

$$\text{FWER} \leq \Pr(\hat{P}_{1:n_0} \leq \alpha_{n-n_0+1}).$$

So, the FWER is strongly controlled at α by a step-down procedure if:

$$\Pr(\hat{P}_{1:n_0} \leq \alpha_{n-n_0+1}) \leq \alpha \qquad \forall \ n_0 = 1, \ldots, n.$$

See, for example, Lehmann and Romano (2005).

The FWER of a step-up procedure satisfies the following:

$$\text{FWER} \leq 1 - \Pr(\hat{P}_{1:n_0} > \alpha_{n-n_0+1}, \ldots, \hat{P}_{n_0:n_0} > \alpha_n).$$

So, the FWER is strongly controlled at α by a step-up procedure if:

$$1 - \Pr(\hat{P}_{1:n_0} > \alpha_{n-n_0+1}, \ldots, \hat{P}_{n_0:n_0} > \alpha_n) \leq \alpha \qquad \forall \ n_0 = 1, \ldots, n.$$

See, for example, **?**.

2.3.1 Bonferroni Procedure

The Bonferroni procedure is considered for the first time by Fisher for pairwise comparisons. It is one of the first used multiple testing procedures to

strongly control the FWER and still being widely used nowadays. Bonferroni procedure is a single-step procedure and could be used to perform general multiple tests besides pairwise comparisons.

The Bonferroni procedure rejects the overall null hypothesis $H_0 : \bigcap_{i=1}^{n} H_i$ if $P_{(1)} \leq \alpha/n$. Furthermore, the test rejects any individual hypothesis H_i if the corresponding P_i is less than α/n.

The Bonferroni Inequality ensures that:

$$P\left\{\bigcup_{i=1}^{n}(P_i \leq \alpha/n)\right\} \leq \sum_{i=1}^{n} \Pr(P_i \leq \alpha/n) \leq \alpha \tag{2.3}$$

Therefore, the probability of rejecting one or more true null hypothesis is less than or equal to α. This procedure actually controls the PFER which is the upper bound of the FWER.

The Bonferroni procedure is very simple to apply, requires no distributional assumption and works under any kind of dependency. It also enables to make decisions on individual hypotheses. However, the procedure is way too conservative and lacks power, especially when highly correlated tests are used. Šidák increases the critical values from α/n to $1-(1-\alpha)^n$ under certain conditions. But the improvement is slight for $\alpha = 0.05$ and $n < 10$.

2.3.2 Holm's Procedure

In Holm (1979), a more powerful stepwise Bonferroni procedure was proposed. While Bonferroni procedure is a single step procedure, Holm's procedure is a step-down procedure based on the closure method. See more details about the closure method in Section 2.1.5.

Holm's procedure rejects $H_{(1)}, \ldots, H_{(j)}$, where

$$j = \max \left\{ k : P_{(i)} < \frac{\alpha}{n-i+1} \text{ for all } i = 1, \ldots, k \right\}$$

It is a step-down procedure and strongly controls the FWER. It works under any kind of dependence structure and is uniformly more powerful than the Bonferroni procedure. It operates as follows. Start from the minimum p-value, $P_{(1)}$. If $P_{(1)} > \alpha/n$, stop and accept all of the hypotheses; otherwise, reject $H_{(1)}$ and continue to compare $P_{(2)}$ with $\alpha/(n-1)$, if greater, accept all the remaining hypotheses; otherwise, reject $H_{(2)}$ and go to $P_{(3)}$ and so on. This procedure is basically applying Bonferroni procedure to every possible subset of the hypotheses. That is, Holm's Procedure is using the single-step Bonferroni procedure as its level-α test to each intersection, thus becomes a short-cut of the closure method as we mentioned in Section 2.1.5.

In Holm's procedure, every $P_{(i)}$ is compared with α_i, which is $\alpha/(n-i+1)$. From the derivation of the critical values, we could see that each critical value is guarding the FWER for the intersection hypotheses of a fixed size, that is, a certain number of hypotheses are jointly true. For example, if $n = 10$, α_3 would be $\alpha/8$ and it would guarantee to control the FWER in the case that 8 out of 10 hypotheses are actually true. Based on this principle, Shaffer (1986) proposed an improvement on Holm's procedure. In Holm's procedure, when the first j p-values are rejected, the $(j+1)$th p-value is compared with $\alpha/(n-j)$. Shaffer suggested using α/j^*, instead of $\alpha/(n-j)$, where j^* is the maximum possible number of true hypotheses given at least j hypotheses are false. In fact, if $j^* \neq n-j$, the corresponding critical value could be lifted up to 1 without losing control of the FWER, which makes the critical values non-monotone. Shaffer's modification ensures the monotonicity of the critical

values.

2.3.3 Hommel's Procedure

As Simes' procedure fails to provide a rule to make decisions over the individual hypothesis when the overall null hypothesis H_0 is rejected, Hommel (1988) employed the closure principle to Simes' procedure and yield a new procedure which controls the FWER strongly.

Hommel's procedure first finds $j = \max\{k \in \{1, \ldots, n\} : P_{(n-k+i)} > i\alpha/k \text{ for } i = 1, \ldots, k\}$ If the maximum does not exist, reject all H_i, otherwise, reject those ones with $P_i \leq \alpha/j$.

Hommel's procedure operates as a two-stage procedure. The first stage is to determine the number of true null hypotheses (n_0). Its estimate of n_0 is the largest size of the intersection hypothesis that can not be rejected by Simes' test. The second stage is to apply the Bonferroni procedure to the p-values with respect to the estimation of n_0 in the first stage.

Hommel's procedure is uniformly more powerful than Holm's procedure. But it only works when Simes' test works.

2.3.4 Hochberg's Procedure

Hochberg (1988) proposed another modification based on Simes' Procedure to strongly control the FWER. It is more powerful than Holm's Procedure too. But, as in Hommel's Procedure, it only works under certain kind of dependence structure.

Hochberg's Procedure rejects all $H_{(1)}, \ldots, H_{(j)}$ where $j = \max\{i \in \{1, \ldots, n\}, P_i < \alpha/(n - i + 1)\}$. If such j does not exist, accept all hypotheses.

Hochberg's procedure is a step-up procedure and strongly controls the FWER of course, whenever Simes' procedure protects the FWER weakly. It uses the same critical values as in Holm's Procedure (See Section 2.3.2), thus it is uniformly more powerful than Holm's procedure as long as it works. It operates as follows. Start from the largest p-value, $P_{(n)}$, if $P_{(n)} < \alpha$, stop and reject all the hypotheses; otherwise, accept $H_{(n)}$ and go on to compare $P_{(n-1)}$ with $\alpha/2$. If smaller, reject all the remaining hypotheses; otherwise, accept $H_{(n-1)}$ and proceed to compare $P_{(n-2)}$ with $\alpha/3$, etc. Hochberg controls the FWER in strong sense due to the property of the closure principle.

Hochberg's procedure is easier to apply than Hommel's procedure. However, Hommel's procedure is uniformly more powerful than Hochberg's as proved in Hommel (1989). Liu (1996) unified Holm's, Hochberg's and Hommel's procedures by a general closed testing procedure to include step-down and step-up procedures by using a critical value matrix. From the matrix, he provided another proof for the advantage of Hommel's procedure over Hochberg's. Also, it is interesting to see that Hochberg's procedure could be presented in the following manner:

Hochberg's procedure first finds $j = \max\{k \in \{1, \ldots, n\} : P_{(n-k+i)} > \alpha/(k-i+1) \text{ for } i = 1, \ldots, k\}$ If the maximum does not exist, reject all H_i, otherwise, reject those ones with $P_i \leq \alpha/j$

In this way, it uses the same setup as Hommel's Procedure except for the critical values at each step. Since $\alpha/(k-i+1) \leq i\alpha/k$ when $i \leq k$. Hochberg's procedure is using smaller critical values at each step, thus is less powerful than Hommel's procedure.

Liu (1996) elaborated the critical matrices for various procedures. But,

his review was heavily based on the assumption of Simes' procedure. He used the word *Dominant* to show the advantage of Hommel's procedure over Holm's. But as we mentioned before, Hommel's procedure does not control FWER strongly under negative dependence while Holm's procedure will control FWER without any assumption on the dependence structure. Therefore, *Dominant* is not valid for all scenarios. Closed testing procedure has two essential ingredients. The first is a size α global test of $H_I = \bigcap_{i \in I} H_i$, for each $I \in K$, where K is the set of all non-empty subsets of $\{1, \ldots, n\}$. The second is the systematic way of making decisions on each H_I: reject H_I if and only if all H_J with $J \supseteq I$, $J \in K$, are rejected by the corresponding tests. In Hochberg and Hommel's Procedure, the size α test is the Simes' test for overall hypothesis. In Holm's procedure, the size α test is the Bonferroni test. Thus, the question still remains: Is there any improvement on Holm's procedure to strongly control the FWER under any kind of dependence structure? Seneta and Chen (2005) put an effort to find more powerful procedures by incorporating correlations. See more details in Section 2.3.8.

2.3.5 Rom's Procedure

As Rom (1990) pointed out, the superiority of Hommel's procedure over Hochberg's is due to the conservatism of Hochberg's procedure, whose size of test is strictly less than α for $n > 2$. This is obvious as it reduces every critical value of Simes', which is an α-level test, for $n > 2$. Under independence assumption, Rom developed a new set of critical values $\{c_{n_n}, \ldots, c_{2_n}, c_{1_n}\}$ to replace $\{\alpha, \frac{\alpha}{2}, \ldots, \frac{\alpha}{n}\}$ from Hochberg's procedure. Please note that with the independence assumption, the p-values are independent uniform (0,1) random

variables.

In order to control the FWER, we need:

$$A_n(\alpha) = \Pr(P_{(n)} > c_{n_n}, P_{(n-1)} > c_{(n-1)_n}, \ldots, P_{(1)} > c_{1_n}) > 1 - \alpha$$

where $\{P_{(1)}, \ldots, P_{(n)}\}$ are the ordered statistics of the independent uniform random variables.

For $n = 1$, it is obvious that $c_{1_1} = \alpha$. Now, suppose we already have $\{c_{1_n}, \ldots, c_{n_n}\}$. When we increase n to $n+1$, by letting $c_{i_n} = c_{(i+1)_{(n+1)}}$ for $1 \leq i \leq n$, we will automatically have all critical values available except for the smallest one, $c_{1_{n+1}}$. And for a general n,

$$A_n(\alpha) = \sum_{i=0}^{n-1} c_{n_n}^i A_1(\alpha) - \sum_{i=1}^{n-1} \binom{n}{i} c_{(n-i)_n}^{n-i} A_i(\alpha) = 1 - \alpha$$

By letting $A_i(\alpha) = 1 - \alpha$, for $1 \leq i \leq n$, we have:

$$\sum_{i=1}^{n-1} c_{n_n}^i - \binom{n}{i} c_{(n-i)_n}^{n-i} = 0 \tag{2.4}$$

Since we already have $\{c_{2_n}, \ldots, c_{n_n}\}$, we can calculate c_{1_n} by (2.4).

However, Rom (1990) failed to consider the monotonicity of the critical values that he derived. The monotonicity of the critical values is essential for the implementation of a step-up multiple testing procedure as stated in Finner and Roters (1998), see also Sarkar (2000). Dalal and Mallows (1992) was able to prove that the critical values in Rom (1990) are monotone, but the result is only valid under the independence assumption. Effort of finding the full generalization of the existence of monotone critical values in more general dependence cases was dashed by Finner and Roters (1998) providing a counterexample, while Sarkar (2000) was able to prove the existence of monotone critical values for $n = 3$ under a restricted dependence structure.

2.3.6 Dunnett and Tamhane's SUP

Dunnett and Tamhane (1992) extended Hochberg's procedure by incorporating the dependence structure of the test statistics and yielded a uniformly more powerful test. Their proof and calculation are based on the normality, equicorrelated and equivariance assumption. This procedure attempts to find a set of critical values satisfying the condition $\{c_1 \leq c_2 \leq \cdots \leq c_n\}$, and accept $H_{(i)}$ if and only if $H_{(j)}, j = 1, \ldots, i-1$ are accepted and $X_{(i)} < c_i$ in terms of test statistics. Therefore, it is a step-up procedure.

As stated in Dunnett and Tamhane (1992), for a step-up procedure to strongly control FWER at level-α, assuming there are n_0 true hypotheses, the following inequality must be followed:

$$\Pr((X_{(1)}, \ldots, X_{(n_0)}) < (c_1, \ldots, c_{n_0})) \geq 1 - \alpha \tag{2.5}$$

where $\{(X_{(1)}, \ldots, X_{(n_0)}) < (c_1, \ldots, c_{n_0})\}$ denotes $\{(X_{(1)} < c_1, \ldots, X_{(n_0)} < c_{n_0}\}$. It is easy to see that when we have the critical values satisfying the equality of (2.5), we will control the FWER at exact α, thus the procedure yields an optimal size α test. c_i's can be sequentially calculated by letting $n_0 = \{1, \ldots, n\}$, similar to the calculation of Rom's critical values, which we stated in terms of p-values. The question still remains for whether these critical values derived are non-decreasing. Dunnett and Tamhane (1992) claimed their optimism by calculating the critical values up to $n = 8$ for dependence case, and $n = 1000$ for independence. Dunnett and Tamhane (1995) proposed a method for calculating the critical values for this procedure with unequally correlated parameter estimates.

2.3.7 Tamhane, Liu and Dunnett's SUDP

Tamhane et al. (1998) proposed a generalized step-up-down procedure and derive a method to calculate the critical values for controlling the FWER with different starting position of r.

The set of critical values is $\{c_1^r, \ldots, c_n^r\}$ which is related to r. The procedure starts from position r with corresponding test statistic $X_{(r)}$ and compare it with c_r^r. If larger, the test will reject $\{H_{(r)}, \ldots, H_{(n)}\}$ and proceed in a step-down manner by testing $X_{(r-1)} > c_{r-1}^r$. If smaller, the test will accept $\{H_{(1)}, \ldots, H_{(r)}\}$ and proceed in a step-up manner by testing $X_{(r+1)} > c_{r+1}^r$.

This procedure is often more appropriate than simply a step-up or step-down procedure and often more powerful. For example, in assessing the superiority of a test drug over placebo and known active controls, it might be necessary to make some preliminary comparisons to see if the clinical trial is sensitive in the sense of being able to detect significant differences between at least a specified number of the known actives and the placebo. A generalized SUDP would be appropriate in this case.

For this procedure to strongly control the FWER at α, the following inequality must be followed:

$$\Pr((X_{(1)}, \ldots, X_{(n)}) < (\underbrace{c_r, \ldots, c_r}_{r}, c_{r+1}, \ldots, c_n)) \geq 1 - \alpha \tag{2.6}$$

(2.6) can be solved recursively by letting $m = r + 1$ and then $m = r + 2$ and so on. Note the first critical value is obviously α. The SUDP reduces to a step-up procedure by choosing $r = 1$ and a step-down procedure by choosing $r = n$. Tamhane et al. (1998) showed this procedure will be optimal when we choose $r = n_0 + 1$. The choice of $r = 1$ achieves nearly the highest power in all

cases, therefore the step-up procedure can always be used without much loss of power.

2.3.8 Seneta and Chen's Procedure

In Section 2.3.4, we raised a question: Is there any improvement on Holm's Procedure to strongly control the FWER under any kind of dependence structure? Seneta and Chen (2005) derived a step-down procedure by utilizing the bivariate distributions of the test statistics. The procedure is uniformly more powerful than Holm's Procedure, and still strongly controls the FWER. The calculations of the critical values need a complicated algorithm, while it is much simpler than Dunnett and Tamhane (1992). The procedure also requires the knowledge of all bivariate correlations of the test statistics.

For a step-down procedure, strong FWER controlling property is presented as:

$$\Pr(\text{at least one true null hypothesis is rejected} \mid n_0 \text{ hypotheses are true}) \leq \alpha \tag{2.7}$$

for $\forall\, n_0 = 1, \ldots, n$. Thus, the critical values in a step-down procedure need to satisfy the inequalities below, see Sarkar (2002a):

$$\Pr(P_{1:n_0} < \alpha_{n-n_0+1}) \leq \alpha \quad \text{for } n_0 = 1, \ldots, n \tag{2.8}$$

There are a total of n inequalities in (2.8) and each of them contains only one α_i. Therefore, each α_i could be calculated independently of each other, from its own inequality.

Due to the property of closure principle, a step-down procedure could be viewed as processing in the following setup:

We first test the intersection hypothesis of size n, i.e., $n_0 = n$ by testing $P_{(1):n} < \alpha_1$ where α_1 satisfies $\Pr(P_{(1):n} < \alpha_1 \mid n_0 = n) \leq \alpha$. If not, we reject $n_0 = n$, hence the smallest p-value and go on to test if $n_0 = n - 1$ by testing $P_{(2):n} < \alpha_2$, where α_2 satisfies $\Pr(P_{(1):n-1} < \alpha_2 \mid n_0 = n - 1) \leq \alpha$. In this setup, each α_i is derived to protect the FWER for a specific $n_0 = n - i + 1$.

Seneta and Chen (2005) define their procedure as follows to satisfy (2.8):

Let:

$$\gamma = \max_{i \in I} \sum_{j \in I - \{i\}} \Pr(P_i \leq \frac{\alpha}{m}, P_j \leq \frac{\alpha}{m} | H_s, s \in I, \text{true})$$

where I is the set of true null hypotheses of fixed size m, $m = 1, \ldots, n$, $\gamma = 0$ when $m = 1$. Then the critical values satisfying (2.8) are given as:

$$\alpha_{n-m+1} = \min\left(\frac{\alpha + \gamma}{m}, \frac{\alpha}{m-1}\right)$$

The monotonicity of the critical values is automatically taken care of by limiting each one to the interval of $[\frac{\alpha}{n-i+1}, \frac{\alpha}{n-i}]$. In this thesis, we will propose another procedure to eliminate this limitation, resulting a more powerful test especially when the test statistics are highly correlated.

Under general dependence structure, Seneta and Chen's procedure could be viewed as using Kounias Inequality to give a smaller upper bound of the probability in (2.8) than Bonferroni's, which was used in Holm's Procedure, thus to improve the critical values. Seneta and Chen (2005) also proposed to apply an even sharper inequality given in Hunter (1976), with increased complexity.

2.4 Procedures controlling the k-FWER

Assume that $\hat{P}_1, \ldots, \hat{P}_{n_0}$ are the p-values corresponding to the true null hypotheses. Then, the k-FWER of a step-down procedure satisfies the following:

$$\text{FWER} \leq \Pr(\hat{P}_{k:n_0} \leq \alpha_{n-n_0+k}).$$

So, the k-FWER is strongly controlled at α by a step-down procedure if:

$$\Pr(\hat{P}_{k:n_0} \leq \alpha_{n-n_0+k}) \leq \alpha \qquad \forall \ n_0 = k, \ldots, n.$$

See, for example, Lehmann and Romano (2005).

The k-FWER of a step-up procedure satisfies the following:

$$\text{FWER} \leq 1 - \Pr(\hat{P}_{k:n_0} > \alpha_{n-n_0+k}, \ldots, \hat{P}_{n_0:n_0} > \alpha_n).$$

So, the k-FWER is strongly controlled at α by a step-up procedure if:

$$1 - \Pr(\hat{P}_{k:n_0} > \alpha_{n-n_0+k}, \ldots, \hat{P}_{n_0:n_0} > \alpha_n) \leq \alpha \qquad \forall \ n_0 = k, \ldots, n.$$

See, for example, **?**.

2.4.1 Korn et al.'s k-FWER procedure

Korn et al. (2004) proposed a step-down permutation-based procedure to control the k-FWER strongly.

Let there be $P_{(1)}, \ldots, P_{(n)}$ ordered p-values calculated from the univariate tests based on n variables, and $H_{(1)}, \ldots, H_{(n)}$ be the corresponding null hypotheses. Their procedure automatically rejects $H_{(1)}, \ldots, H_{(k-1)}$ and test $P_{(k)} > y^{\alpha}_{k,k}$. If greater, the procedure accepts the current and all the remaining hypotheses and stop. If not, the procedure rejects $H_{(k)}$ and go on to test

$P_{(k+1)} > y^{\alpha}_{k+1,k}$, where $y^{\alpha}_{i,k}$ are computed based on multivariate permutation distribution under the global null hypothesis, i.e., all variables satisfy the null hypothesis. The closure principle ensures this procedure control the k-FWER at level-α. However, the monotonicity of the critical values has been violated in this procedure with $y^{\alpha}_{i,k} = 1$ for $i = 1, \ldots, k-1$.

2.4.2 Lehmann and Romano's k-FWER procedure

Lehmann and Romano (2005) proposed another step-down procedure to control the k-FWER at level-α with critical values defined in (2.9):

$$\alpha_i = \begin{cases} \frac{k\alpha}{n} & i = 1, \ldots, k-1 \\ \frac{k\alpha}{n-i+k} & i = k, \ldots, n \end{cases} \tag{2.9}$$

Lehmann and Romano's procedure can be viewed as a modified Holm's procedure with control of the k-FWER instead of the FWER. Lehmann and Romano (2005) proved their procedure to control the k-FWER strongly under no assumption on the dependence structure and the procedure is sharp in the sense that there exists a joint distribution for the p-values to exactly control the k-FWER at α. Of course under certain dependence assumption, the procedure can be improved, see Section 2.4.3.

Lehmann and Romano (2005) also mentioned that one could always reject the hypotheses corresponding to the smallest $k-1$ p-values. But for the sake of monotonicity, they kept their critical values the same as α_k.

Since Lehmann and Romano's procedure is a modified Holm's procedure, question still remains for whether there exists an improvement like Seneta and Chen's procedure over Holm's. This thesis will try to give a similar improvement in Chapter 4.

2.4.3 Sarkar's procedures

Inspired by the idea of generalizing FWER to k-FWER, Sarkar (2008a) proposed a corresponding generalization of Simes' test to allow rejection of the intersection once at least k of the null hypotheses are rejected, instead of the original Simes' procedure using $k = 1$. Then through utilizing the kth order joint distribution of p-values, Sarkar (2008a) proposed a generalized Hochberg procedure by using his generalized Simes' test.

Sarkar's Generalized Hochberg procedure's critical values $\{\alpha_k, \ldots, \alpha_n\}$ are given by:

$$G_k(\alpha_i) = \frac{k(k-1)\cdots 1}{(n-i+k)(n-i+k-1)\cdots(n-i+1)}\alpha, \qquad i = k, \ldots, n$$

where G_k is the common cdf of the maximum of any k of the n P_i's under the null hypotheses, that is,

$$G_k(\alpha_i) = \Pr(P_1 < \alpha_i, \ldots, P_k < \alpha_i)$$

assuming that the p-values have an identical kth order joint null distributions.

Sarkar's procedure also kept the first $k-1$ critical values to be the same as α_k. Under simulation, the procedure is shown to be more powerful than Lehmann and Romano's procedure in step-up manner when the p-values are close to being independent. In this thesis, we will propose another procedure to be more powerful than Lehmann and Romano's procedure when the p-values are highly correlated.

2.5 A Comparison between Step-Down and Step-Up procedures

In this section, we will present a brief comparison between step-down and step-up procedures.

Generally, with the same critical values, step-up procedures are more powerful than step-down procedures as they reject at least as many as the step-down procedures do. However, this is a lame comparison as mentioned in Finner and Roters (1998). Holm's procedure always controls the FWER while Hochberg's procedure may fail to do so. If we develop a step-down and a step-up procedure both controlling the FWER at exact α, their comparison will become more complicated, that is, step-down procedures will have larger critical values. This can already be seen when we have independent p-values for multiple testing. Having Šidák's procedure in the step-down manner and Rom's values in the step-up manner, both procedures control the FWER at exact α, but it is known that Šidák has larger critical values than Rom. Actually, this property is quite obvious and holds for all scenarios. When a step-down procedure is an exact level-α test, its corresponding step-up procedure will always exceed the level-α, since it is more powerful than the step-down procedure. To keep it under α, one has to reduce the critical values for a step-up procedure. Therefore, the preference of the step-up procedure is in doubt. As Finner and Roters (1998) proved that the difference between step-down critical values and step-up critical values asymptotically goes to 0 if n goes to infinity, the probability of step-up procedures having more rejections than step-down procedures should go to 0 too.

Unlike in step-up procedures, where one must derive their critical values in a recursive manner and the calculation becomes extremely cumbersome when n is large, in step-down procedures the critical values are derived independently of each other. Note that (2.8) contains n inequalities, with each inequality involving only one α_i. Another advantage for step-down procedures is the similarity between these inequalities in terms of mathematical manipulations, which greatly simplify the calculations of the critical values. Now let's consider a specific $n_0 = k$ in (2.8):

$$\Pr(P_{(1):k} < \alpha_{n-k+1}) = \Pr\left(\bigcup_{j=1}^{k}(P_j < \alpha_{n-k+1})\right) \leq \alpha \tag{2.10}$$

From (2.10), $\Pr\left(\bigcup_{j=1}^{k}(P_j < \alpha_{n-k+1}) \mid \text{All } k \text{ hypotheses are true}\right)$ represents a union of events and there are several upper bounds that have been developed in literature. Each upper bound give a step-down procedure to control the FWER.

Table 2.1 shows the current best available FWER controlling procedures under each kind of dependence structure and each type of multiple testing procedure.

Table 2.1: Current available procedures to control the FWER strongly

	All dependency Unknown Correlation	All dependency Known Correlation	PRDS	Independent
Single Step	Bonferroni	Bonferroni	Bonferroni	Sidak
Step-Down	Holm	Seneta and Chen	Senata and Chen	Sidak
Step-Up	None	Dunnett and Tamhane (small n) Hochberg (large n)	Dunnett and Tamhane (small n) Hochberg (large n)	Rom
Two-Stage	None	None	Hommel	Hommel

CHAPTER 3

IMPROVED HOLM'S PROCEDURES UNDER DEPENDENCE

In this chapter, we propose new step-down procedures controlling the FWER using the bivariate correlations of the underlying test statistics. The critical values of the procedures are uniformly larger than the original Holm's procedure under any kind of dependency as long as the bivariate correlations are available. Simulations are carried out to compare the new procedure with Holm's and other existing FWER controlling procedures under the same condition.

3.1 An Alternative Global Test Using Correlations

In this section, we will first present a method, an alternative to Bonferroni, for testing an intersection hypothesis using correlations. The critical value in this method is calculated by obtaining an explicit formula for the Type-I error rate of a single-step test for an intersection hypothesis in terms of its critical value and the underlying correlations and equating it to α. This method is used to test each subset of hypotheses while applying the closed testing principle, thereby yielding a step-down procedure different from Holm's with strong FWER control.

More specifically, let us consider testing the intersection of a subset of n_0 hypotheses by a test where each p-value is compared with a common critical value $\alpha(n_0) = \alpha_{n-n_0+1}$. We will reject the intersection hypothesis if at least one p-value is less than $\alpha(n_0)$. The Type-I error rate of this method is:

$$\begin{aligned} \text{Type-I error} = &\ \Pr(R \geq 1) \\ = &\ \sum_{i=1}^{n_0} \Pr(R = i) \end{aligned} \tag{3.1}$$

where R is the number of p-values less than $\alpha(n_0)$. We will determine $\alpha(n_0)$ so that the above Type-I error rate is controlled at α.

Towards deriving an explicit formula for (3.1) in terms of $\alpha(n_0)$, we will present the following two lemmas by splitting (3.1) into two parts. Let us consider, without any loss of generality, the case when $n_0 = n$, and $\alpha(n) = \alpha_c$.

Lemma 3.1

$$\begin{aligned} & Pr(R=1) \\ = & \sum_{i=1}^{n} Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=2}^{n} \frac{1}{r-1} Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c) \end{aligned}$$

where $P_{n-1:n-2}^{(-i,-j)} = 1$.

Lemma 3.2

$$\begin{aligned} & \sum_{r=2}^{n} Pr(R=r) \\ = & \sum_{r=2}^{n} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{r(r-1)} Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c) \end{aligned}$$

Proof of Lemma 3.1 :

$$\begin{aligned} & \Pr(R=1) \\ = & \sum_{i=1}^{n} \Pr(P_i < \alpha_c, P_{1:n} < \alpha_c, P_{2:n} > \alpha_c) \end{aligned} \tag{3.2}$$

By taking P_i out of the order statistics:

$$(3.2) = \sum_{i=1}^{n} \Pr(P_i < \alpha_c, P_{1:n-1}^{(-i)} > \alpha_c) \tag{3.3}$$

By $\Pr(A \text{ and } B) = \Pr(A) - \Pr(A \text{ and } B^c)$:

$$\begin{aligned} (3.3) = & \sum_{i=1}^{n} \Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \Pr(P_i < \alpha_c, [P_{1:n-1}^{(-i)} > \alpha_c]^c) \end{aligned} \tag{3.4}$$

By expanding $[P_{1:n-1}^{(-i)} > \alpha_c]^c$:

$$\begin{aligned}(3.4) = \quad & \sum_{i=1}^{n} \Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \Pr(P_i < \alpha_c, P_{1:n-1}^{(-i)} < \alpha_c, P_{2:n-1}^{(-i)} > \alpha_c) \\ & - \sum_{i=1}^{n} \Pr(P_i < \alpha_c, P_{2:n-1}^{(-i)} < \alpha_c, P_{3:n-1}^{(-i)} > \alpha_c) \\ & - \ldots - \sum_{i=1}^{n} \Pr(P_i < \alpha_c, P_{n-1:n-1}^{(-i)} < \alpha_c)\end{aligned} \tag{3.5}$$

By taking P_j out of the order statistics:

$$\begin{aligned}(3.5) = \quad & \sum_{i=1}^{n} \Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{1:n-2}^{(-i,-j)} > \alpha_c, P_{2:n-2}^{(-i,-j)} > \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{2} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{1:n-2}^{(-i,-j)} < \alpha_c, P_{2:n-2}^{(-i,-j)} > \alpha_c) \\ & - \ldots - \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{n-1} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{n-2:n-2}^{(-i,-j)} < \alpha_c) \\ = \quad & \sum_{i=1}^{n} \Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=2}^{n} \frac{1}{r-1} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c)\end{aligned}$$

Proof of Lemma 3.2:

$$\begin{aligned} & \sum_{r=2}^{n} \Pr(R = r) \\ = \quad & \sum_{r=2}^{n} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r:n} < \alpha_c, P_{r+1:n} > \alpha_c)\end{aligned} \tag{3.6}$$

By taking P_i and P_j out of the order statistics:

$$(3.6) = \sum_{r=2}^{n} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{r(r-1)} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c)$$

Based on the above two lemmas, we are now ready to present the following theorem.

Theorem 3.1 *For a single-step test with the critical value α_c for testing an intersection hypothesis of size n,*

$$\begin{aligned} \textit{Type-I error} = & \sum_{i=1}^{n} Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=2}^{n} \frac{1}{r} Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c) \end{aligned}$$

Proof. of Theorem 3.1:

$$\begin{aligned} & \text{Type-I error} \\ = \; & \Pr(R = 1) + \sum_{r=2}^{n} \Pr(R = r) \\ = \; & \sum_{i=1}^{n} \Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=2}^{n} \frac{1}{r-1} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c) \\ & + \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=2}^{n} \frac{1}{r(r-1)} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c) \\ = \; & \sum_{i=1}^{n} \Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=2}^{n} \frac{1}{r} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c) \end{aligned} \tag{3.7}$$

Theorem 3.1 provides us an explicit form of the Type-I error rate, when we apply a single-step test. If we could solve the equation obtained by equating (3.7) to α, the corresponding test would be an optimal test in that it would be the least conservative single-step test for the intersection hypothesis. However, the equation is usually too complicated to solve if there are correlations

involved. Often, Holm-Type step-down procedures turn to look for a compromised way to trade optimality for simplicity. That is, they find an upper bound for (3.7) and let the upper bound equal to α. For example, Holm's procedure sacrifices the whole second term in (3.7) for ultra simplicity, as seen in the next corollary. Also, note that Holm-Type step-down procedures require monotone critical values. Sarkar (2002a) noted that if we can solve the critical values directly from (3.7), the critical values are automatically monotone. However, for upper bound solutions, the monotonicity of the critical values is not obvious. We will discuss more about monotonicity in Section 3.2.

Corollary 3.1 *Holm's procedure controls the FWER strongly.*

Proof. In Holm's procedure, his critical values are decided by the following equation:

$$\sum_{i=1}^{n} \Pr(P_i < \alpha_c) = \alpha \tag{3.8}$$

As the p-values follow uniform distribution under null hypothesis, (3.8) becomes:

$$n\alpha_c = \alpha \quad \Rightarrow \quad \alpha_c = \alpha/n$$

Similarly, for the other n_0's, $\alpha_k = \alpha/n_0$, where $k = n - n_0 + 1$. This set of α_k's can be used as tests for all intersection hypotheses and the resulting procedure controls the FWER strongly due to closed testing.

Obviously, we could find smaller upper bound, $U(\alpha_k, n_0)$ to improve Bonferroni and Holm's procedure by making use of the second term in (3.7). Note there exists certain distributions to make the second term of (3.7) equal to 0, hence Holm's procedure is sharp in some occasions, see Lehmann and Romano

(2005) for an example. Nevertheless, in most scenarios, we could utilize the second term to improve Holm's procedure.

We will now present an improved form of the Type-I error rate in the following theorem by making use of the bivariate correlations between p-values. Without loss of generality, again we will only prove for the case of $n_0 = n$:

Theorem 3.2 *The Type-I error rate of a single-step test for testing an intersection hypothesis of size n based on a critical value α_c is given by:*

$$\begin{aligned} \textit{Type-I error} = \quad & n\alpha_c - \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{n} Pr(P_i < \alpha_c, P_j < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=3}^{n} \frac{1}{r(r-1)} Pr(P_i < \alpha_c, P_j < \alpha_c, R_{n-2} \leq r-3) \end{aligned} \tag{3.9}$$

where R_{n-2} is the number of rejections based on $n-2$ tests excluding i^{th} and j^{th} tests, that is, $\{R_{n-2} = r\}$ denotes $\{P^{(-i,-j)}_{r:n-2} < \alpha_c, P^{(-i,-j)}_{r+1:n-2} > \alpha_c\}$ and $\{R_{n-2} \leq r\}$ denotes $\{P^{(-i,-j)}_{r+1:n-2} > \alpha_c\}$.

Proof of Theorem 3.2: Let's start from (3.7),

$$
\begin{aligned}
& \text{Type-I error} \\
= \; & n\alpha_c - \sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=2}^{n}\frac{1}{r}\Pr(P_i<\alpha_c, P_j<\alpha_c, P^{(-i,-j)}_{r-2:n-2}<\alpha_c, P^{(-i,-j)}_{r-1:n-2}>\alpha_c) \\
= \; & n\alpha_c - \sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=2}^{n}\frac{1}{r}\Pr(P_i<\alpha_c, P_j<\alpha_c, R_{n-2}=r-2) \\
= \; & n\alpha_c - \sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=2}^{n}\frac{1}{r}\Pr(P_i<\alpha_c, P_j<\alpha_c, R_{n-2}\leq r-2) \\
& \quad + \sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=3}^{n}\frac{1}{r}\Pr(P_i<\alpha_c, P_j<\alpha_c, R_{n-2}\leq r-3) \\
= \; & n\alpha_c - \sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=3}^{n+1}\frac{1}{r-1}\Pr(P_i<\alpha_c, P_j<\alpha_c, R_{n-2}\leq r-3) \\
& \quad + \sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=3}^{n}\frac{1}{r}\Pr(P_i<\alpha_c, P_j<\alpha_c, R_{n-2}\leq r-3) \\
= \; & n\alpha_c - \sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{n}\Pr(P_i<\alpha_c, P_j<\alpha_c, R_{n-2}\leq n-2) \\
& \quad - \sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=3}^{n}\frac{1}{r(r-1)}\Pr(P_i<\alpha_c, P_j<\alpha_c, R_{n-2}\leq r-3) \\
= \; & n\alpha_c - \sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{n}\Pr(P_i<\alpha_c, P_j<\alpha_c) \\
& \quad - \sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=3}^{n}\frac{1}{r(r-1)}\Pr(P_i<\alpha_c, P_j<\alpha_c, R_{n-2}\leq r-3)
\end{aligned}
\tag{3.10}
$$

Theorem 3.2 provides us another explicit form of the Type-I error rate by separating bivariate distributions of the p-values from higher dimension distributions. We may then take advantage of this form and improve the Bonferroni procedure.

Remark 3.1 *The solutions of α_k in the following equations control the Type-I*

error rate at α for any intersection hypothesis of size $(n-k+1)$.

$$U(\alpha_k, n_0) = \sum_{i=1}^{n_0} Pr(P_i < \alpha_k) - \sum_{i=1}^{n_0} \sum_{j \neq i} \frac{1}{n_0} Pr(P_i < \alpha_k, P_j < \alpha_k) = \alpha$$

Hence we find an improved test for any intersection hypothesis and it could be used in a step-down manner to create an improved Holm's procedure as long as the derived critical values are monotone. As we do not achieve the sharp case in this test, it remains unknown for the monotonicity of the critical values we get. We will make an effort to prove the monotonicity property in Section 3.2.

3.2 Monotonicity of the Critical Values in Step-Down Procedures

In this section, we will attack the problem raised in the previous section - the monotonicity of the critical values derived in the test we derived in Section 3.1. First, we will present the following proposition.

Proposition 3.1 *When an upper bound, $U(\alpha_k, n_0)$, has the following property: It is non-decreasing by α_k and non-decreasing by n_0, the critical values derived from Remark 3.1 will be monotone.*

Proof. We consider consecutive upper bounds $U(\alpha_k, n_0)$ and $U(\alpha_{k+1}, n_0 - 1)$. Assume we have solved α_k from equation $U(\alpha_k, n_0) = \alpha$, as U is a non-decreasing function by n_0, inequality $U(\alpha_k, n_0 - 1) \leq U(\alpha_k, n_0)$ holds. In order to achieve $U(\alpha_{k+1}, n_0 - 1) = \alpha$, we need to have $\alpha_{k+1} \geq \alpha_k$ since U is a non-decreasing function in α_k. This completes the proof.

Lemma 3.3 *Let the random variables X and Y be jointly distributed as a standard bivariate normal. Then the probability of $Pr(X < t, Y > t)$ is non-increasing in t, when $t > 0$.*

Proof. For bivariate standard normal distribution,

$$\Pr(X > t, Y < t) = \int_t^\infty \Phi\left(\frac{t - y\rho}{\sqrt{1-\rho^2}}\right)\varphi(y)d_y$$

Take derivative w.r.t. t, the above becomes:

$$\begin{aligned}
&-\Phi\left(\frac{t - t\rho}{\sqrt{1-\rho^2}}\right)\varphi(t) + \int_t^\infty \varphi\left(\frac{t - y\rho}{\sqrt{1-\rho^2}}\right)\varphi(y)\frac{1}{\sqrt{1-\rho^2}}d_y \\
\leq &-\frac{1}{2}\varphi(t) + \int_t^\infty \varphi\left(\frac{t - y\rho}{\sqrt{1-\rho^2}}\right)\varphi(y)\frac{1}{\sqrt{1-\rho^2}}d_y \\
= &-\frac{1}{2}\varphi(t) + \int_t^\infty \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}e^{-\frac{(t-y\rho)^2}{2(1-\rho^2}}\frac{1}{2\pi}e^{-\frac{y^2}{2}}d_y \\
= &-\frac{1}{2}\varphi(t) + \int_t^\infty \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}e^{-\frac{y^2-2yt\rho+t^2}{2(1-\rho^2)}}\frac{1}{\sqrt{2\pi}}d_y \\
= &-\frac{1}{2}\varphi(t) + \int_t^\infty \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}e^{-\frac{(y-\rho t)^2}{2(1-\rho^2}}d_y\frac{1}{2\pi}e^{-\frac{t^2}{2}} \\
\leq &-\frac{1}{2}\varphi(t) + \frac{1}{2}\varphi(t) = 0
\end{aligned}$$

The last step follows from $\Pr(Y > t) \leq \Pr(Y > \rho t) = \frac{1}{2}$, when $t > 0$. Therefore, we prove that the derivative is always non-positive, which means the probability is always non-increasing when $t > 0$. This means $\Pr(X > t) - \Pr(X > t, Y > t)$ is always non-increasing, no matter what the correlation is, as long as $t > 0$. Lemma 3.3 could be easily transformed to p-values with $\alpha < 0.5$.

Theorem 3.3 *The critical values calculated in Remark 3.1 and in Seneta and Chen (2005) are monotone, given the test statistics are normally distributed and $\alpha < 0.5$.*

Proof. Our derived upper bounds in Remark 3.1 are:

$$\begin{aligned} U(\alpha_k, n_0) &= \sum_{i=1}^{n_0} \Pr(P_i < \alpha_k) - \sum_{i=1}^{n_0} \sum_{j \neq i} \frac{1}{n_0} \Pr(P_i < \alpha_k, P_j < \alpha_k) \\ &= \sum_{i=1}^{n_0} \left(\Pr(P_i < \alpha_k) - \sum_{j \neq i} \frac{1}{n_0} \Pr(P_i < \alpha_k, P_j < \alpha_k) \right) \\ &= \sum_{i=1}^{n_0} \sum_{j \neq i} \left(\frac{1}{n_0 - 1} \Pr(P_i < \alpha_k) - \frac{1}{n_0} \Pr(P_i < \alpha_k, P_j < \alpha_k) \right) \\ &= \sum_{i=1}^{n_0} \sum_{j \neq i} \left(\frac{1}{n_0(n_0 - 1)} \Pr(P_i < \alpha_k) + \frac{1}{n_0} \Pr(P_i < \alpha_k, P_j > \alpha_k) \right) \end{aligned} \tag{3.11}$$

By the result from Lemma 3.3, both terms in (3.11) are non-decreasing functions in α_k, where $k = n - n_0 + 1$. It is obvious that $U(\alpha_k, n_0)$ is a non-decreasing function by n_0 by taking the difference of $U(\alpha_k, n_0)$ and $U(\alpha_k, n_0 + 1)$.

3.3 Improved Holm's Procedure Using Correlations

We will present our new step-down procedure:

Theorem 3.4 *A step-down procedure using the critical values calculated from the equations below controls the FWER at* α.

$$U(\alpha_k, n_0) = \sum_{i=1}^{n_0} Pr(P_i < \alpha_k) - \sum_{i=1}^{n_0} \sum_{j \neq i} \frac{1}{n_0} Pr(P_i < \alpha_k, P_j < \alpha_k) = \alpha \tag{3.12}$$

where $n_0 = 1, \ldots, n$, $k = n - n_0 + 1$.

Proof. This result is obvious by Theorem 3.2 and the closure method. Theorem 3.2 provides us a level-α test for any intersection hypotheses and the

step-down procedure ensures us a systematic way to apply the test to every possible intersection hypotheses, hence it controls the FWER at α.

It is interesting to note that Seneta and Chen (2005) provides a step-down procedure which is very similar to what we derived here. Their derived critical values are as follows:

Seneta and Chen's Procedure: For any given subset hypotheses of size n_0, let:

$$\gamma_k = \max_{i=1,\ldots,n_0} \sum_{j \neq i}^{n_0} \Pr(P_i \leq \frac{\alpha}{n_0}, P_j \leq \frac{\alpha}{n_0})$$

The critical value α_k is derived as:

$$\alpha_k = \min\left(\frac{\alpha + \gamma_k}{n_0}, \frac{\alpha}{n_0 - 1}\right)$$

Their corresponding upper bound is:

$$U(\alpha_k, n_0) = \max\left\{n_0\alpha_k - \max_{i=1,\ldots,n_0} \sum_{j \neq i}^{n_0} \Pr(P_i < \alpha/n_0, P_j < \alpha/n_0), (n_0 - 1)\alpha_k\right\}$$

while our upper bound is:

$$U(\alpha_k, n_0) = n_0\alpha_k - \sum_{i=1}^{n_0} \sum_{j \neq i} \frac{1}{n_0} \Pr(P_i < \alpha_k, P_j < \alpha_k)$$

There are several differences between our new procedure and CS procedure as referred to in Seneta and Chen (2005).

1. Regarding the bivariate terms, we are using an "average" type of summations while CS uses a "max" type, which means their procedure has the advantage over ours, but we can prove that our procedure could be using a "max" type upper bound to erase this disadvantage. Detailed proof is given in Corollary 3.2. Note that under the assumption that the

null hypotheses have identical distribution and common correlation, our procedure and CS procedure will be the same at this point.

2. CS procedure presented an explicit form to find α_k. They derived this form by sacrificing the exact cut-off points and replacing them with conservative ones (by changing α_k to α/n_0). Critical values would be simpler to calculate but the complicity of using the exact cut-off points is not a major problem for modern computers. Detailed algorithm is given below.

3. CS procedure has one more restriction, which is restraining the critical values to be less than the next corresponding Holm's critical value, that is, $\alpha_k \leq \frac{\alpha}{n_0-1}$. In this way, as we mentioned before, the critical values keep the monotonicity property automatically. In this thesis, we showed that this might be a redundant restriction, at least for normal statistics, see more details in Section 3.2.

4. Our procedure gives out an explicit form of the Type-I error rate of the global test and we can easily detect the conservativeness of the developed step-down procedures by evaluating the third term in (3.10). We also specifically paved a way for further improvements by utilizing the third term in (3.10), if we are allowed to bring in higher dimension of correlation. Seneta and Chen's procedure does not have such ability.

Corollary 3.2 *The solution of α_c in the following equations control the Type-I error rate at α for any intersection hypothesis of size n.*

$$U(\alpha_c, n) = n\alpha_c - \max_{j=1,\ldots,n} \sum_{i \neq j} Pr(P_i < \alpha_c, P_j < \alpha_c) = \alpha \tag{3.13}$$

Proof. Although this corollary is a direct interpretation of the Kounias Inequality, see Kounias (1968), we will show the exact conservativeness of the Kounias Inequality in our proof.

We will start from (3.7):

$$\begin{aligned}(3.7) = \quad & \sum_{i=1}^{n} \Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{n} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{n-2:n-2}^{(-i,-j)} < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=2}^{n-1} \frac{1}{r} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c)\end{aligned} \tag{3.14}$$

By putting P_i and P_j back into the order statistics in the second term:

$$\begin{aligned}(3.14) = \quad & n\alpha_c - (n-1)\Pr(P_{n:n} < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=2}^{n-1} \frac{1}{r} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c)\end{aligned} \tag{3.15}$$

By taking a pre-specified P_k out of the order statistics in the second term first and then taking P_i out:

$$\begin{aligned}(3.15) = \quad & n\alpha_c - \sum_{i=1, i \neq k}^{n} \Pr(P_i < \alpha_c, P_k < \alpha_c, P_{n-2:n-2}^{(-i,-k)} < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=2}^{n-1} \frac{1}{r} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c)\end{aligned} \tag{3.16}$$

The equation should be valid for any pre-specified k, thus for the best bound, we take the particular P_k out where k maximizes $\sum_{i=1, i \neq k}^{n} \Pr(P_i <$

$\alpha_c, P_k < \alpha_c)$, $k = 1, \ldots, n$.

$$\begin{aligned}
(3.16) = \quad & n\alpha_c - \max_{k=1,\ldots,n} \sum_{i=1, i\neq k}^{n} \Pr(P_i < \alpha_c, P_k < \alpha_c) \\
& + \sum_{i=1, i\neq k}^{n} \Pr(P_i < \alpha_c, P_k < \alpha_c, P_{n-2:n-2}^{(-i,-k)} > \alpha_c) \\
& - \sum_{i=1}^{n} \sum_{j\neq i} \sum_{r=2}^{n-1} \frac{1}{r} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c) \\
= \quad & n\alpha_c - \max_{k=1,\ldots,n} \sum_{i=1, i\neq k}^{n} \Pr(P_i < \alpha_c, P_k < \alpha_c) \\
& + \sum_{i=1, i\neq k}^{n} \sum_{r=2}^{n-1} \Pr(P_i < \alpha_c, P_k < \alpha_c, P_{r-2:n-2}^{(-i,-k)} < \alpha_c, P_{r-1:n-2}^{(-i,-k)} > \alpha_c) \\
& - \sum_{i=1}^{n} \sum_{j\neq i} \sum_{r=2}^{n-1} \frac{1}{r} \Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c)
\end{aligned} \tag{3.17}$$

By putting P_i, P_k back into the order statistics in the third term and P_i, P_j back into the order statistics in the fourth term and combine them:

$$\begin{aligned}
(3.17) = \quad & n\alpha_c - \max_{k=1,\ldots,n} \sum_{i=1, i\neq k}^{n} \Pr(P_i < \alpha_c, P_k < \alpha_c) \\
& - \sum_{r=3}^{n-1} (r-2) \Pr(P_{r:n} < \alpha_c, P_{r+1:n} > \alpha_c)
\end{aligned} \tag{3.18}$$

As the third term in (3.18) is always non-positive, solving α_c from:

$$U(\alpha_c, n) = n\alpha_c - \max_{k=1,\ldots,n} \sum_{i=1, i\neq k}^{n} \Pr(P_i < \alpha_c, P_k < \alpha_c) = \alpha$$

and other α_i from similar equations will provide a level-α test for each intersection hypothesis. Thus the resulting step-down procedure will control the FWER strongly. This completes the proof for Corollary 3.2.

Now we will present the algorithm to calculate our critical values. The computation can be done by R and given the assumption of common correla-

tion, the computation time is within seconds for n up to 100, and still within a minute for n as large as 10,000.

Algorithm for the New Procedure: For finding α_k:

1. Find:

$$\gamma_k = \max_{i=1,\ldots,n_0} \sum_{j \neq i} \Pr(P_i \leq \frac{\alpha}{n_0}, P_j \leq \frac{\alpha}{n_0})$$

2. Let $\alpha_{k_1} = \frac{\alpha+\gamma_k}{n_0}$.

3. Find:

$$\gamma_k^{'} = \max_{i=1,\ldots,n_0} \sum_{j \neq i}^{n_0} \Pr(P_i \leq \alpha_{k_1}, P_j \leq \alpha_{k_1})$$

4. Let $\alpha_{k_2} = (\alpha + \gamma_k^{'})/n$ and replace α_{k_1} with it in Step 3.

5. Repeat Step 3 and 4 until the critical value converges.

3.4 Simulations Results and Tables

In this section, we will present the simulations results to show the advantage of our step-down procedure over Holm (1979) and Seneta and Chen (2005).

We first generated 8 independent/dependent normal random variables $N(\mu_i, 1)$ $i = 1, \ldots, 8$, with a common correlation ρ where n_1 (number of false null hypotheses) of these 8 μ_i's are all equal to 2 and the rest are all equal to 0. We then applied to this data set each of the aforementioned procedures to test if each of these means is either true ($\mu_i = 0$) or false ($\mu_i > 0$), and noted what proportion of the n_0 (n_1) means that are all equal to 0 (2) were correctly declared as true (false). We repeated this experiment 20,000 times and

obtained the average of these proportions to obtain the simulated FWER (Average Power) for each procedure. Figure 3.1 compares the committed FWER and Figure 3.2 compares the average powers of these procedures, the six panels presenting this comparison for $\rho = 0, 0.3, 0.5, 0.75, 0.9$ and 0.99.

Table 3.1 summarizes the critical values being used under different occasions. It uses the same setup as the Table 4 in Seneta and Chen (2005). The CS critical values are presented in the parentheses. It could be seen that our critical values are uniformly larger than CS. The higher correlation, the bigger advantage we have.

Remark 3.2 *The critical values for CS are different from the Table 4 in Seneta and Chen (2005) as we are using one-sided tests. Seneta and Chen (2005) was using two-sided tests. The corresponding CS critical values for one-sided tests are calculated by our algorithm.*

Table 3.1: Critical Values for Multivariate Normal Distribution with common correlation ρ and $n = 8$

i	1	2	3	4	5	6	7	8
Holm	0.00625	0.00714	0.00833	0.01000	0.01250	0.01667	0.02500	0.05
$\rho = 0$	0.00628	0.00719	0.00839	0.01008	0.01262	0.01686	0.02532	0.05
	(0.00628)	(0.00719)	(0.00839)	(0.01008)	(0.01262)	(0.01685)	(0.02531)	(0.05)
0.3	0.00650	0.00744	0.00871	0.01048	0.01314	0.01757	0.02628	0.05
	(0.00648)	(0.00742)	(0.00868)	(0.01045)	(0.01309)	(0.01750)	(0.02619)	(0.05)
0.5	0.00693	0.00795	0.00931	0.01121	0.01406	0.01873	0.02766	0.05
	(0.00684)	(0.00784)	(0.00917)	(0.01104)	(0.01382)	(0.01842)	(0.02731)	(0.05)
0.6	0.00735	0.00843	0.00987	0.01188	0.01486	0.01970	0.02874	0.05
	(0.00714)	(0.00818)	(0.00957)	(0.01150)	(0.01438)	(0.01911)	(0.02811)	(0.05)
0.9	0.01209	0.01371	0.01578	0.01853	0.02231	0.02775	0.03612	0.05
	(0.00714)	(0.00833)	(0.01000)	(0.01250)	(0.01667)	(0.02299)	(0.03241)	(0.05)
0.95	0.01572	0.01760	0.01994	0.02292	0.02681	0.03205	0.03936	0.05
	(0.00714)	(0.00833)	(0.01000)	(0.01250)	(0.01667)	(0.02434)	(0.03386)	(0.05)
0.99	0.02611	0.02818	0.03055	0.03330	0.03649	0.04024	0.04469	0.05
	(0.00714)	(0.00833)	(0.01000)	(0.01250)	(0.01667)	(0.02500)	(0.03586)	(0.05)
1.00	0.05000	0.05000	0.05000	0.05000	0.05000	0.05000	0.05000	0.05
	(0.00714)	(0.00833)	(0.01000)	(0.01250)	(0.01667)	(0.02500)	(0.03750)	(0.05)

We next compare our procedure with Seneta and Chen's procedure using a biology example appeared in their paper. The example concerns the

effect of ethanol on sleep time in a sample of 20 rats originally considered in Hattan and Eacho (1978). There are 4 treatments, corresponding to different concentrations of ethanol, each applied to 5 rats. Thus, it is a multivariate t-distribution with $\nu = 16$. Treatment 1 is distilled water (control group) and treatments 2, 3, 4 are increasing concentrations of ethanol. We then use the pairwise comparisons to determine if there is any difference between different treatments. A total of 6 hypotheses being tested are: $H^{(i,j)} : \mu_i - \mu_j = 0, i < j$. Their derived p-values are listed as follows: $1.021 \times 10^{-6}, 9.819 \times 10^{-5}, 2.312 \times 10^{-4}, 0.01025, 0.02435$ and 0.04011. Bonferroni procedure would reject the first 3 hypotheses and Tukey's test would reject the first 4 hypotheses. Holm's procedure, which would reject all 6 hypotheses, has a borderline case for the second largest p-value, with 0.02435 being compared to the critical value of 0.025. Seneta and Chen improved this critical value to 0.273 and our new procedure further improved it to 0.276. Interested readers could refer to Seneta and Chen (2005) for more details.

3.5 Some Remarks

We have offered in this chapter a step-down procedure to control the FWER under any type of dependence structure as long as the first-order dependence between the test statistics expressed through the bivariate probabilities can be enumerated. As numerically demonstrated for normally distributed test statistics, our proposed method utilizing a knowledge of the correlations performs much better compared to both Holm's and Seneta-Chen's procedures. However, it is important to note that our procedure could be improved and be generalized as well. We will remark on several possible improvements in the

following.

Remark 3.3 *In many applications, particularly if n is large, one might be willing to tolerate more than one false rejection provided the number of such cases is controlled, thereby increasing the ability of the procedure to detect false null hypotheses. This suggests replacing control of the FWER by controlling the probability of k or more false rejections, which we call the k-FWER. Procedures controlling k-FWER are discussed in Lehmann and Romano (2005), see also Korn et al. (2004) and Romano and Wolf (2005). As k-FWER is a generalization of the notion of the FWER, one should be able to generalize procedures controlling the FWER to those controlling the k-FWER. Lehmann and Romano (2005) gave such a k-FWER procedure generalizing Holm's procedure. Our approach in the context of the FWER can be easily generalized to that in the context of k-FWER by changing $Pr(R \geq 1)$ to $Pr(R \geq k)$ in (3.1).*

Remark 3.4 *Shaffer (1986) proposed a modified Holm's procedure when the free combinations condition does not hold. From our derivation of the critical values, it is easy to see why there exists such an improvement and we could modify our critical values in the same way.*

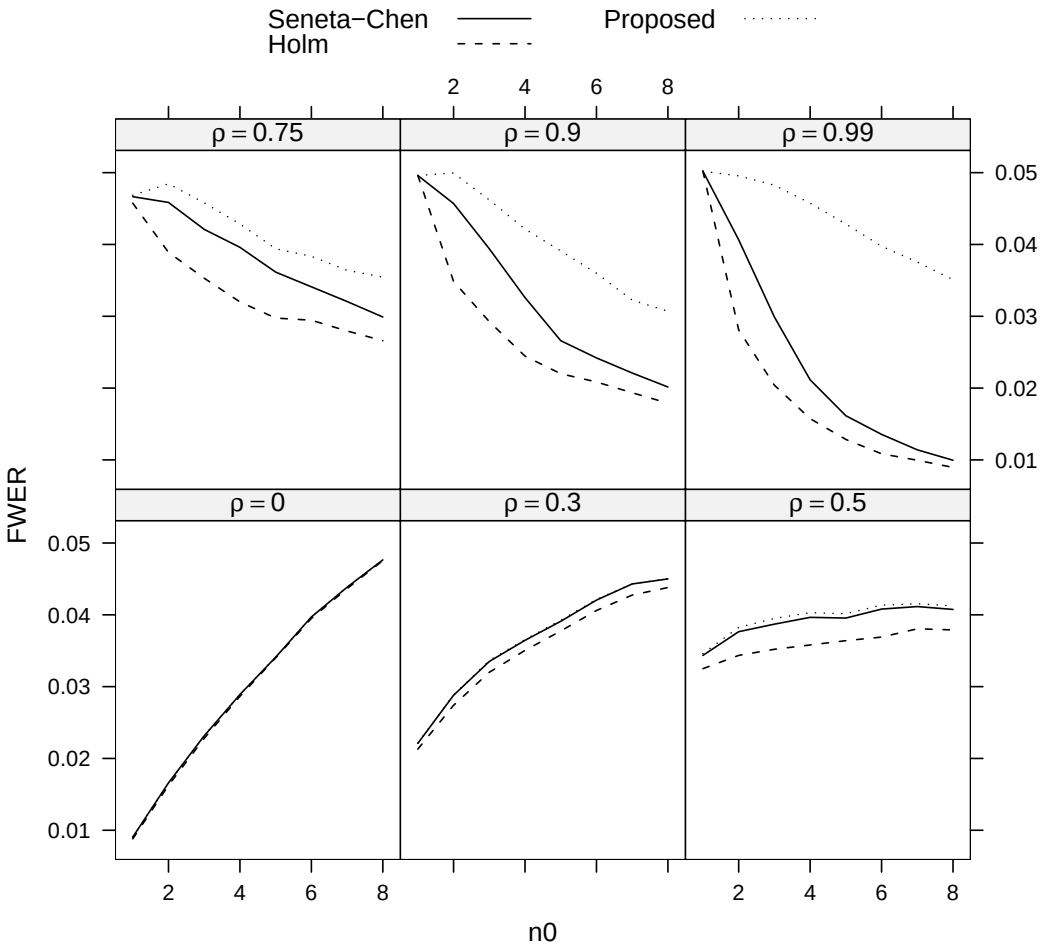

Figure 3.1: Comparison of actual FWER of the proposed procedure, CS procedure and the original Holm's procedure for equicorrelated multivariate normal with $\alpha = 0.05$

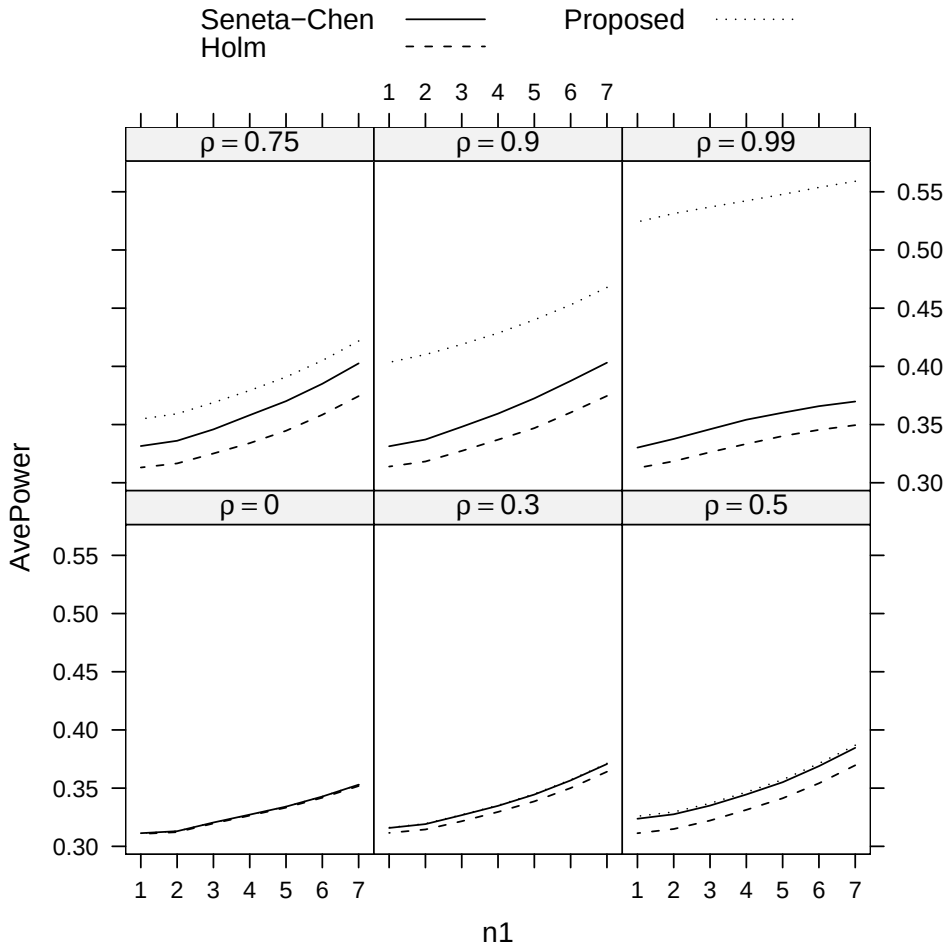

Figure 3.2: Comparison of powers of the proposed procedure, CS procedure and the original Holm's procedure for equicorrelated multivariate normal with $\alpha = 0.05$

CHAPTER 4

IMPROVED LEHMAN AND ROMANO'S PROCEDURE

In this chapter, we propose a new step-down procedure controlling the k-FWER using the bivariate correlations of the underlying test statistics. This procedure is proposed in response of Remark 3.3 in Chapter 3. The critical values of the procedures are uniformly larger than Lehman and Romano's procedure under any kind of dependency as long as the bivariate correlations are available. Simulations are carried out to compare the new procedure with Lehman and Romano's procedure.

4.1 A Generalized Global Test Using Correlations

In this section, we will present a method for testing an intersection hypothesis using correlations. The critical value in this method is calculated by

obtaining an explicit formula for the generalized Type-I error rate of a single-step test for an intersection hypothesis in terms of its critical value and the underlying correlations and equating it to α. We define the generalized Type-I error rate to be the probability of having at least k false rejections instead of 1 in the previous chapter. This method is used to test each subset of hypotheses while applying the closed testing principle, thereby yielding a step-down procedure different from Lehman and Romano's with strong k-FWER control.

More specifically, let us consider testing the intersection of a subset of n_0 hypotheses by a test where each p-value is compared with a common critical value $\alpha(n_0) = \alpha_{n-n_0+1}$. We will reject the intersection hypothesis if at least k p-values are less than $\alpha(n_0)$. The Type-I error rate of this method is:

$$\begin{aligned} \text{Type-I error} = &\ \Pr(R \geq k) \\ = &\ \sum_{i=k}^{n_0} \Pr(R = i) \end{aligned} \tag{4.1}$$

where R is the number of p-values less than $\alpha(n_0)$. We will determine $\alpha(n_0)$ so that the above Type-I error rate is controlled at α. We set Type-I error rate to 0 if $n_0 < k$.

Similar to what we derived in Chapter 3, Lemma 3.1 and Lemma 3.2 become:

Lemma 4.1

$$\begin{aligned} & Pr(R = k) \\ = &\ \sum_{i=1}^{n} \frac{1}{k} Pr(P_i < \alpha_c) \\ & - \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r=k+1}^{n} \frac{1}{r-1} Pr(P_i < \alpha_c, P_j < \alpha_c, P_{r-2:n-2}^{(-i,-j)} < \alpha_c, P_{r-1:n-2}^{(-i,-j)} > \alpha_c) \end{aligned}$$

where $P_{n-1:n-2}^{(-i,-j)} = 1$.

Lemma 4.2

$$\begin{aligned}&\sum_{r=k+1}^{n} Pr(R=r)\\ =\ &\sum_{r=k+1}^{n}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{r(r-1)}Pr(P_i<\alpha_c, P_j<\alpha_c, P_{r-2:n-2}^{(-i,-j)}<\alpha_c, P_{r-1:n-2}^{(-i,-j)}>\alpha_c)\end{aligned}$$

Proof of Lemma 4.1 and 4.2: The proofs are similar to those in Chapter 3.

Based on the above two lemmas, we are now ready to present the following theorem.

Theorem 4.1 *For a single-step test with the critical value α_c for testing an intersection hypothesis of size n,*

$$\begin{aligned}\textit{Type-I error} = &\sum_{i=1}^{n}\frac{1}{k}Pr(P_i<\alpha_c)\\ &-\sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=k+1}^{n}\frac{1}{r}Pr(P_i<\alpha_c, P_j<\alpha_c, P_{r-2:n-2}^{(-i,-j)}<\alpha_c, P_{r-1:n-2}^{(-i,-j)}>\alpha_c)\end{aligned}$$

Proof. of Theorem 4.1:

$$\begin{aligned}&\text{Type-I error}\\ =\ &\Pr(R=k)+\sum_{r=k+1}^{n}\Pr(R=r)\\ =\ &\sum_{i=1}^{n}\frac{1}{k}\Pr(P_i<\alpha_c)\\ &-\sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=k+1}^{n}\frac{1}{r-1}\Pr(P_i<\alpha_c, P_j<\alpha_c, P_{r-2:n-2}^{(-i,-j)}<\alpha_c, P_{r-1:n-2}^{(-i,-j)}>\alpha_c)\\ &+\sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=k+1}^{n}\frac{1}{r(r-1)}\Pr(P_i<\alpha_c, P_j<\alpha_c, P_{r-2:n-2}^{(-i,-j)}<\alpha_c, P_{r-1:n-2}^{(-i,-j)}>\alpha_c)\\ =\ &\sum_{i=1}^{n}\frac{1}{k}\Pr(P_i<\alpha_c)\\ &-\sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=k+1}^{n}\frac{1}{r}\Pr(P_i<\alpha_c, P_j<\alpha_c, P_{r-2:n-2}^{(-i,-j)}<\alpha_c, P_{r-1:n-2}^{(-i,-j)}>\alpha_c)\end{aligned} \tag{4.2}$$

Corollary 4.1 *Lehman and Romano's procedure (2005) controls the k-FWER strongly.*

Proof. In Lehman and Romano's procedure, their critical values are decided by the following equation:

$$\sum_{i=1}^{n} \frac{1}{k} \Pr(P_i < \alpha_c) = \alpha \tag{4.3}$$

As the p-values follow uniform distribution under null hypothesis, (4.3) becomes:

$$\frac{n}{k}\alpha_c = \alpha \quad \Rightarrow \quad \alpha_c = k\alpha/n$$

Similarly, for the other n_0's, $\alpha_i = k\alpha/n_0$, where $i = n - n_0 + 1$. This set of α_i's can be used as tests for all intersection hypotheses and the resulting procedure controls the k-FWER strongly due to closed testing.

Similar to Chapter 3, we will present an improved form of the generalized Type-I error rate in the following theorem by making use of the bivariate correlations between p-values. Without loss of generality, again we will only prove for the case of $n_0 = n$:

Theorem 4.2 *The generalized Type-I error rate of a single-step test for testing an intersection hypothesis of size n based on a critical value α_c is given by:*

$$\begin{aligned} \textit{Type-I error} = \ & \tfrac{n}{k}\alpha_c - \sum_{i=1}^{n}\sum_{j\neq i} \frac{1}{n} Pr(P_i < \alpha_c, P_j < \alpha_c) \\ & - \sum_{i=1}^{n}\sum_{j\neq i}\sum_{r=k+2}^{n} \frac{1}{r(r-1)} Pr(P_i < \alpha_c, P_j < \alpha_c, R_{n-2} \leq r-3) \end{aligned} \tag{4.4}$$

where R_{n-2} is the number of rejections based on $n-2$ tests excluding i^{th} and j^{th} tests, that is, $\{R_{n-2} = r\}$ denotes $\{P_{r:n-2}^{(-i,-j)} < \alpha_c, P_{r+1:n-2}^{(-i,-j)} > \alpha_c\}$ and $\{R_{n-2} \leq r\}$ denotes $\{P_{r+1:n-2}^{(-i,-j)} > \alpha_c\}$.

Proof of Theorem 4.2: The proof is also similar to the proof of Theorem 3.2.

Theorem 4.2 provides us another explicit form of the generalized Type-I error rate by separating bivariate distributions of the p-values from higher dimension distributions.

Remark 4.1 *The solutions of α_k in the following equations control the generalized Type-I error rate at α for any intersection hypothesis of size $(n-k+1)$.*

$$U(\alpha_k, n_0) = \sum_{i=1}^{n_0} \frac{1}{k} Pr(P_i < \alpha_k) - \sum_{i=1}^{n_0} \sum_{j \neq i} \frac{1}{n_0} Pr(P_i < \alpha_k, P_j < \alpha_k) = \alpha$$

Hence we find an improved test for any intersection hypothesis and it could be used in a step-down manner to create an improved Lehman and Romano's procedure.

4.2 Improved Lehman and Romano's Procedure Using Correlations

We will present our new step-down procedure:

Theorem 4.3 *A step-down procedure using the critical values calculated from the equations below controls the k-FWER at α.*

$$U(\alpha_l, n_0) = \sum_{i=1}^{n_0} \frac{1}{k} Pr(P_i < \alpha_l) - \sum_{i=1}^{n_0} \sum_{j \neq i} \frac{1}{n_0} Pr(P_i < \alpha_l, P_j < \alpha_l) = \alpha \quad (4.5)$$

where $n_0 = 1, \ldots, n$, $l = n - n_0 + 1$.

Proof. This result is obvious by Theorem 4.2 and the closure method. Theorem 4.2 provides us a level-α test for any intersection hypotheses and the step-down procedure ensures us a systematic way to apply the test to every possible intersection hypotheses, hence it controls the k-FWER at α.

Now we will present the algorithm to calculate our critical values.

Algorithm for the New Procedure: For finding α_k:

1. Find:

$$\gamma_k = \max_{i=1,\ldots,n_0} \sum_{j \neq i} \Pr(P_i \leq \frac{\alpha}{n_0}, P_j \leq \frac{\alpha}{n_0})$$

2. Let $\alpha_{k_1} = \frac{k(\alpha+\gamma_k)}{n_0}$.

3. Find:

$$\gamma_k^{'} = \max_{i=1,\ldots,n_0} \sum_{j \neq i}^{n_0} \Pr(P_i \leq \alpha_{k_1}, P_j \leq \alpha_{k_1})$$

4. Let $\alpha_{k_2} = k(\alpha + \gamma_k^{'})/n$ and replace α_{k_1} with it in Step 3.

5. Repeat Step 3 and 4 until the critical value converges.

4.3 Simulations Results and Tables

In this section, we will present the simulations results to show the advantage of our step-down procedure over Lehmann and Romano (2005).

We first generated 10 independent/dependent normal random variables $N(\mu_i, 1)$ $i = 1, \ldots, 10$, with a common correlation ρ where n_1 of these 10 μ_i's are all equal to 2 and the rest are all equal to 0. We then applied to this data set each of the aforementioned procedures to test if each of these means

is either true ($\mu_i = 0$) or false ($\mu_i > 0$), and noted what proportion of the n_0 (n_1) means that are all equal to 0 (2) were correctly declared as true (false). We set $k = 2$ and then repeated this experiment 20,000 times and obtained the average of these proportions to obtain the simulated k-FWER (Average Power) for each procedure. Figure 4.1 compares the committed k-FWER and Figure 4.2 compares the average powers of these procedures, the six panels presenting this comparison for $\rho = 0, 0.25, 0.5, 0.75, 0.9$ and 0.99.

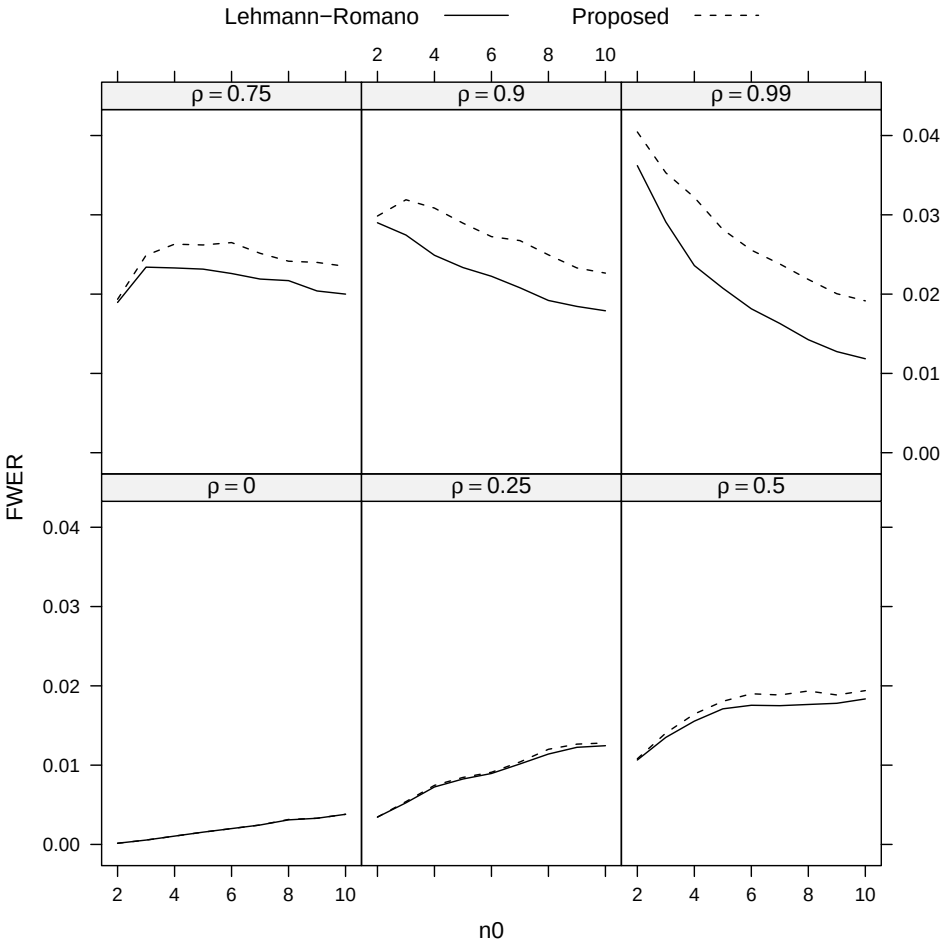

Figure 4.1: Comparison of actual 2-FWER of the proposed procedure and LR procedure for equicorrelated multivariate normal with $\alpha = 0.05$

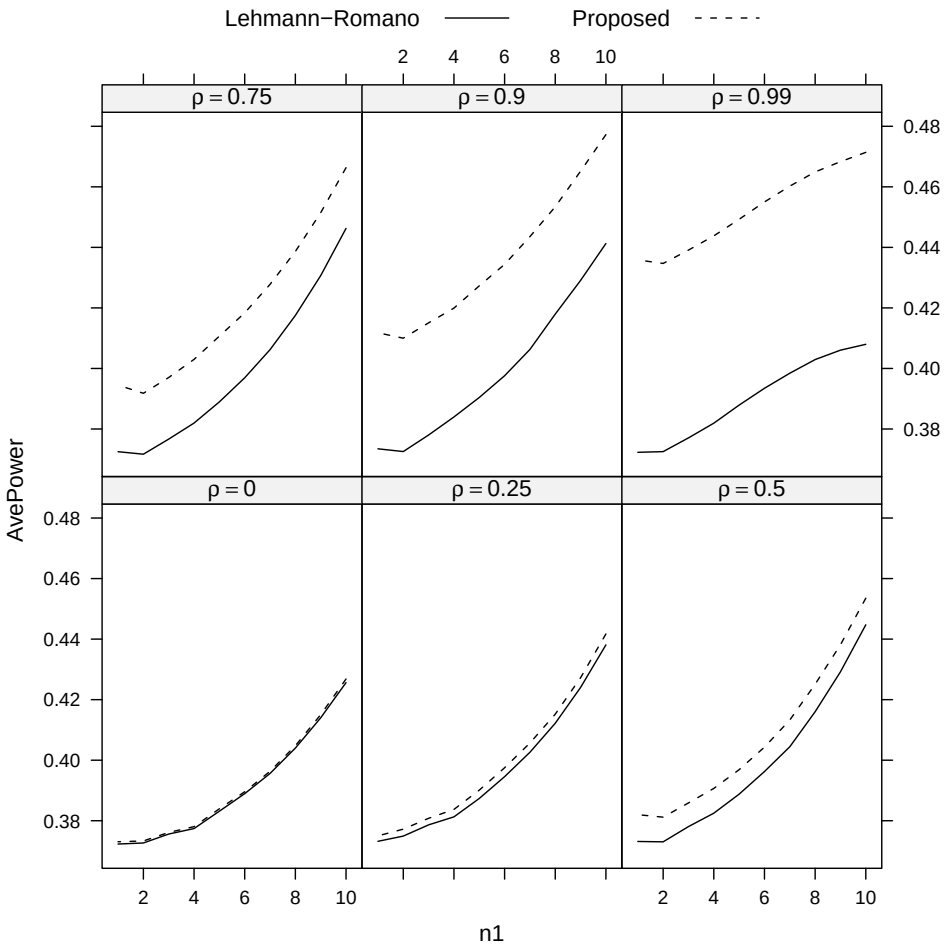

Figure 4.2: Comparison of powers of the proposed procedure and LR procedure for equicorrelated multivariate normal with $\alpha = 0.05$

CHAPTER 5

ANOTHER PROCEDURE TO IMPROVE HOLM'S PROCEDURE UNDER INDEPENDENCE OR WEAK DEPENDENCE

From the simulation results in Chapter 3 and 4, we can see that our newly derived step-down procedures have the least advantage against Holm's procedure under independence or weak dependence. Therefore, we now propose a simple improvement solely for this situation.

5.1 New Procedures to Control the FWER and Generalized FWER under Independence

These procedures are inspired by Storey et al. (2004), which is a procedure controlling the false discovery rate. Benjamini and Hochberg (1995) introduced the false discovery rate. Their critical values, $i\alpha/n$, will control the FDR when used as a step-up test. However, when some of the null hypotheses are not true, the procedure is conservative by a factor π_0 which is the proportion of the true null hypotheses among all null hypotheses. A number of adaptive procedures have also recently been introduced that control the FDR by conservatively estimating pi_0. For instance, Storey (2002) and Storey et al. (2004) proposed a method to estimate the proportion of true null hypotheses, π_0. This quantity, in turn, could be used to improve BH procedure, by raising the critical values from $i\alpha/n$ to $i\alpha/n\hat{\pi}_0$.

Storey's estimation of π_0:

$$\hat{\pi}_0 = \frac{W(\lambda) + 1}{(1-\lambda)n}$$

where λ is a pre-specified number between 0 and 1, $W(\lambda)$ is the number of p-values exceeding λ and n is the number of hypotheses to be tested.

Storey et al. (2004) proved their procedure will control the FDR under independence and some types of *weak dependence*, such as dependence in finite blocks, ergodic dependence and certain mixing distributions.

The similar improvement might also be achieved for FWER procedures. We will first use Storey's method to estimate n_0 from the complete set of

hypotheses and derive a similar approach to improve Holm's procedure. We will show the new procedure in the following conjecture.

New Step-Down FWER Procedure:

For testing n hypotheses, $H_1, H_2, \ldots, H_n$, and their corresponding p-values, $P_1, P_2, \ldots, P_n$, we can use Storey's method to estimate n_0 and derive the following critical values:

$$\alpha_i = \begin{cases} \frac{\alpha}{\hat{n}_0} & \text{if } i \leq n - \hat{n}_0 + 1 \\ \frac{\alpha}{n-i+1} & \text{otherwise.} \end{cases}$$

This set of critical values, when used in a step-down manner, will control the FWER strongly under independence.

Based on the procedure above, we are also proposing the following similar improvement for Lehman and Romano's k-FWER procedure.

New Step-Down k-FWER Procedure:

For testing n hypotheses, $H_1, H_2, \ldots, H_n$, and their corresponding p-values, $P_1, P_2, \ldots, P_n$, we can use Storey's method to estimate n_0 and derive the following critical values:

$$\alpha_i = \begin{cases} \frac{k\alpha}{\hat{n}_0} & \text{if } i \leq n - \hat{n}_0 + 1 \\ \frac{k\alpha}{n-i+1} & \text{otherwise.} \end{cases}$$

This set of critical values, when used in a step-down manner, will control the k-FWER strongly under independence.

Simulation results show the control of FWER or k-FWER for both of the above procedures.

5.2 Simulation Results

We first generated 100 independent normal random variables $N(\mu_i, 1)$ $i = 1, \ldots, 100$ where n_1 of these 100 μ_i's are all equal to 2 and the rest are all equal to 0. We then applied to this data set the procedure for independence to test if each of these means is either true ($\mu_i = 0$) or false ($\mu_i > 0$), and noted what proportion of the n_0 (n_1) means that are all equal to 0 (2) were correctly declared as true (false). We repeated this experiment 20,000 times and obtained the average of these proportions to obtain the simulated FWER (Average Power). Figure 5.1 compares the committed FWER and Figure 5.2 compares the average powers for our new procedure and Holm's procedure.

Then we generated 10 independent normal random variables $N(\mu_i, 1)$ $i = 1, \ldots, 10$ where n_1 of these 10 μ_i's are all equal to 2 and the rest are all equal to 0. We then applied to this data set the procedure for independence to test if each of these means is either true ($\mu_i = 0$) or false ($\mu_i > 0$), and noted what proportion of the n_0 (n_1) means that are all equal to 0 (2) were correctly declared as true (false). We set $k = 2$. We repeated this experiment 20,000 times and obtained the average of these proportions to obtain the simulated 2-FWER (Average Power). Figure 5.3 compares the committed 2-FWER and Figure 5.4 compares the average powers for our new procedure and Holm's procedure.

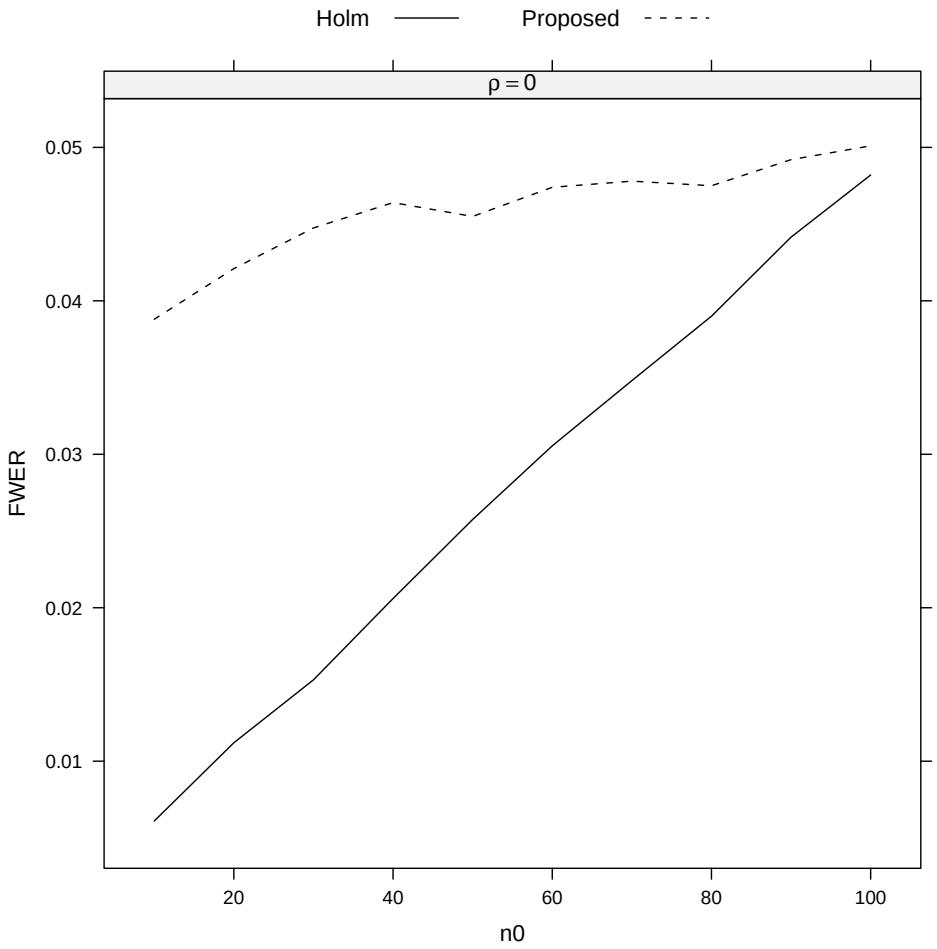

Figure 5.1: Comparison of actual FWER of the proposed procedure and the original Holm's procedure for independent multivariate normal with $\alpha = 0.05$

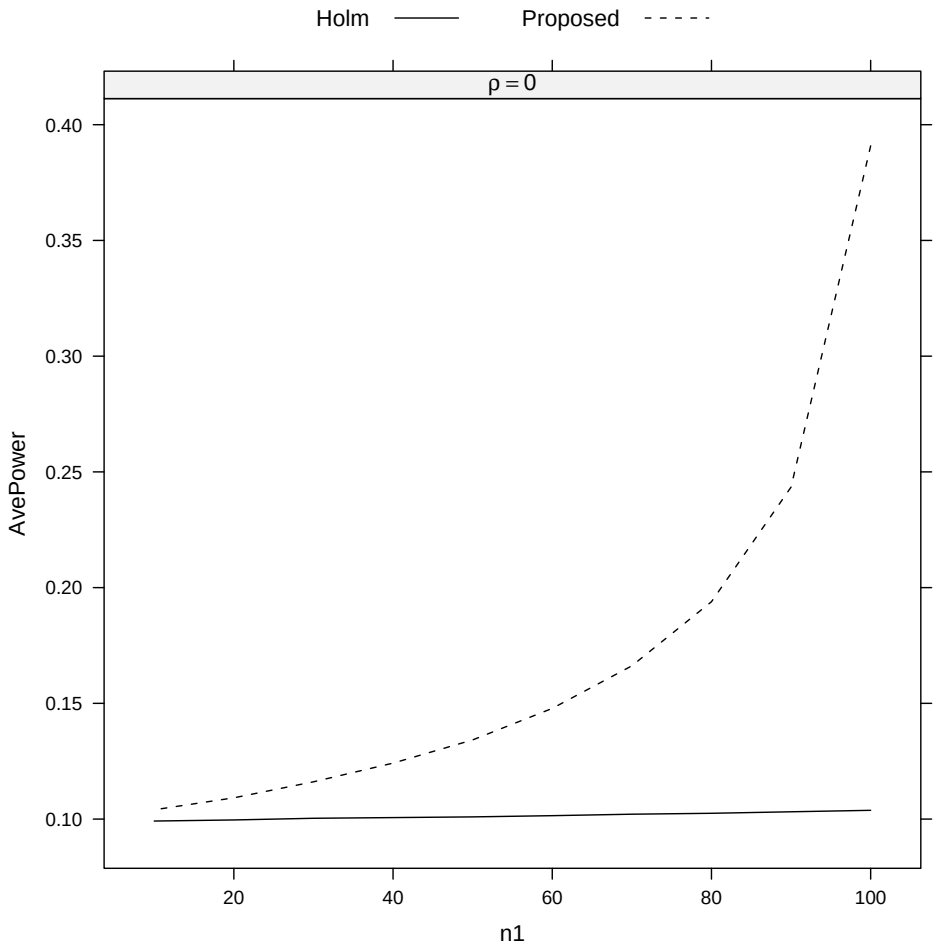

Figure 5.2: Comparison of powers of the proposed procedure and the original Holm's procedure for independent multivariate normal with $\alpha = 0.05$

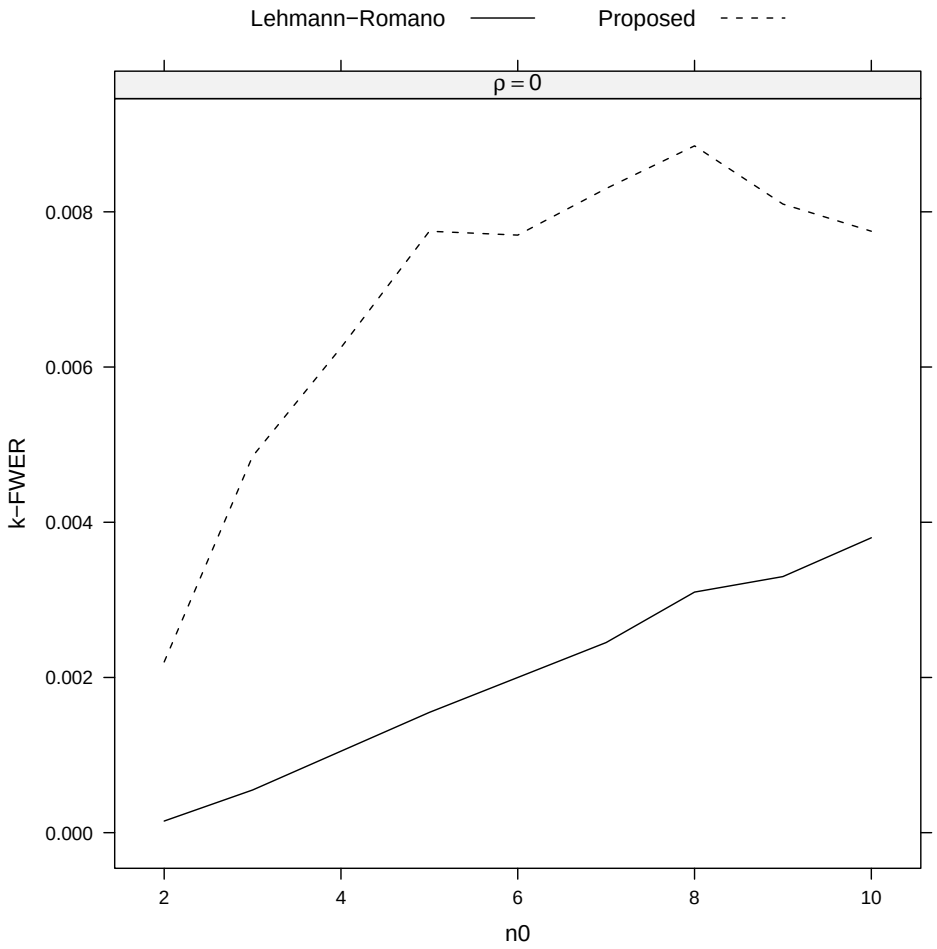

Figure 5.3: Comparison of actual 2-FWER of the proposed procedure and the LR procedure for independent multivariate normal with $\alpha = 0.05$

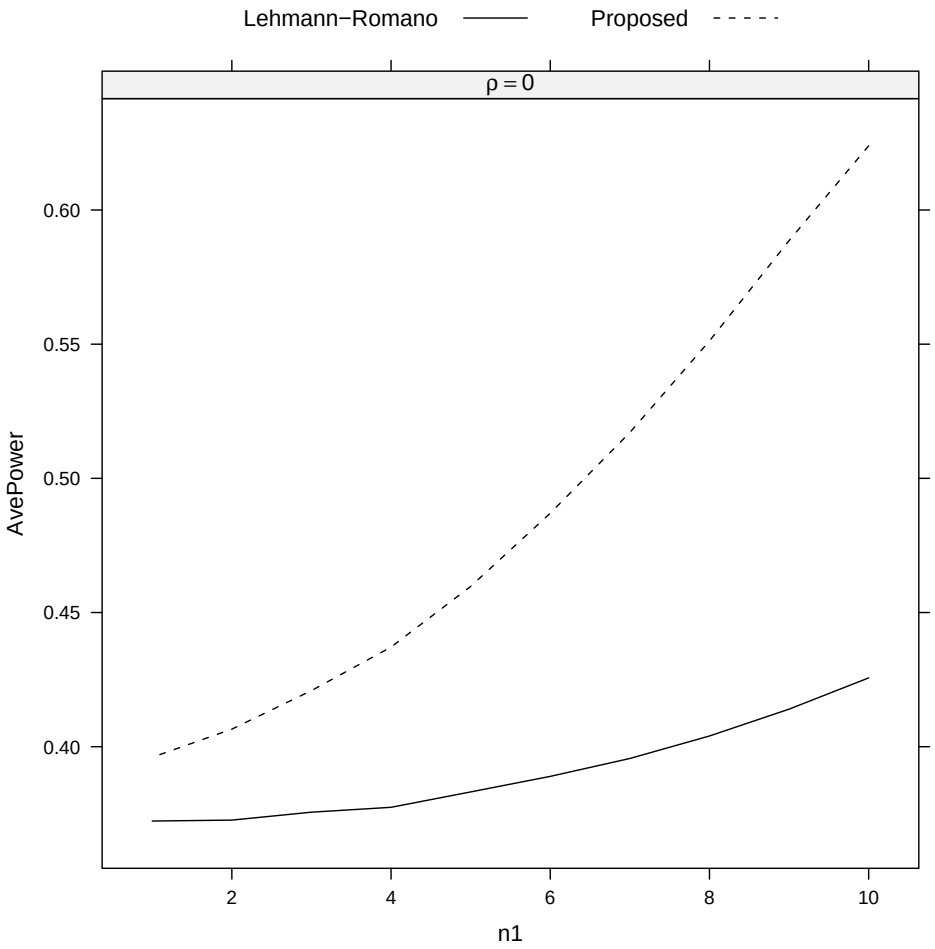

Figure 5.4: Comparison of powers of the proposed procedure and the LR procedure for independent multivariate normal with $\alpha = 0.05$

CHAPTER 6

FUTURE RESEARCH

In the previous chapters, we developed Step-Down procedures to control the FWER or k-FWER. Our results were based on two earlier theories: The closure method and the Kounias Inequality. Also, we proposed two procedures to control FWER and k-FWER under independence. In the future, we will try to prove their controllability. We will also try to extend the estimation of n_0 to dependence cases and it could then be combined with our new procedures proposed in Chapter 3 and 4. Moreover, future research will be done to try to prove in general that the critical values derived from our "upper bound" approach will be automatically monotone.

REFERENCES

Amaratunga, D. and J. Cabrera (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. Hoboken, New Jersey: John Wiley & Sons, Inc.

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

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

Cai, G. and S. K. Sarkar (2006). Modified simes' critical values under positive dependence. *Journal of Statistical Planning and Inference 136*, 4129–4146.

Cai, G. and S. K. Sarkar (2008). Modified simes' critical values under independence. *Statistics and Probability Letters 78*(12), 1362–1368.

Dalal, S. and C. Mallows (1992). Buying with exact confidence. *Annals of Applied Probablity 2*, 752–765.

Dunnett, C. W. and A. Tamhane (1995). Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics 51*, 217–227.

Dunnett, C. W. and A. C. Tamhane (1992). A step-up multiple test procedure. *Journal of the American Statistical Association 87*, 162–170.

Finner, H. and M. Roters (1998). Asymptotic comparisons of step-up and step-down multiple test procedures based on exchangeable test statistics. *Annals of Statistics 26*, 505–520.

Gabriel, K. (1969). Simultaneous test proceduressome theory of multiple comparisons. *Annals of Mathematical Statistics 40*, 224–250.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B 64*, 499–517.

Grinold, R. C. and R. N. Kahn (2000). *Active Portfolio Management* (Second ed.). New York: McGraw-Hill.

Hattan, D. and P. Eacho (1978). Relationship of ethanol blood level to rem and non-rem sleep time and distribution in the rat. *Life Sciences 22*, 839–846.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika 75*, 800–802.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple Comparison Procedures*. John Wiley & Sons.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified test procedure. *Biometrika 75*, 383–386.

Hommel, G. (1989). A comparison of two modified bonferroni prodedures. *Biometrika 76*, 624–625.

Hunter, D. (1976). An upper bound for the probability of a union. *Journal of Applied Probability 13*, 597–603.

Korn, E., J. Troendle, L. Mcshane, and R. Simon (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference 124*, 379–398.

Kounias, E. (1968). Bounds for the probability of a union of events, with applications. *Annals of Mathematical Statistics 39*, 2154–2158.

Kwong, K. S., B. Holland, and S. H. Cheung (2002). A modified benjamini-hochberg multiple comparisons procedure for controlling the false discovery rate. *Journal of Statistical Planning and Inference 104*, 351–365.

Lehmann, E. and J. Romano (2005). Generalizations of the familywise error rate. *Annals of Statistics 33*, 1138–1154.

Liu, W. (1996). Multiple tests of a non-hierarchical finite family of hypotheses. *Journal of the Royal Statistical Society Series B (Methodological) 58*(2), 455–461.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*, 655–660.

Miller, R. (1966). *Simultaneous Statistical Inference*. McGraw-Hill.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified bonferroni inequality. *Biometrika 77*, 663–665.

Romano, J. P. and M. Wolf (2005). Control of generalized error rates in multiple testing. Technical Report, Department of Statistics, Stanford University.

Roy, S. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics 24*, 220–238.

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of the simes conjecture. *Annals of Statistics 26*(2), 494–504.

Sarkar, S. K. (2000). A note on the monotonicity of the critical values of a step-up test. *Journal of Statistical Planning and Inference 87*, 241–249.

Sarkar, S. K. (2002a). Recent advances in multiple hypothesis testing. *Journal of Statistical Studies special volume in honor of Prof. M.M. Ali's 65th birthday*, 293–306.

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

Sarkar, S. K. (2004). Fdr-controlling stepwise procedures and their false negatives rates. *Journal of Statistical Planning and Inference 125*, 119–137.

Sarkar, S. K. (2007). Stepup procedures controlling generalized FWER and generalized FDR. *Annals of Statistics 35*(6), 2405–2420.

Sarkar, S. K. (2008a). Generalizing simes' test and hochberg's stepup procedure. *Annals of Statistics 36*(1), 337–363.

Sarkar, S. K. (2008b). Two stage step-up procedures controlling FDR. *Journal of Statistical Planning and Inference 138*(4).

Sarkar, S. K. and C.-K. Chang (1997). The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association 92*(440), 1601–1608.

Sarkar, T. K. (1969). Some lower bounds of reliability. Technical Report, Operation Research and Statistics, Stanford University.

Seneta, E. and J. T. Chen (2005). Simple stepwise tests of hypotheses and multiple comparisons. *International Statistical Review 73*, 21–34.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association 81*, 826–831.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

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

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics 31*(6), 2013–2035.

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

Tamhane, A., W. Liu, and C. W. Dunnett (1998). A generalized step-up-down multiple test procedures. *Canadian Journal of Statistics 26*, 353–363.