UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Solving Consistency Problems in Multiple Hypotheses Testing with Consonant Likelihood Ratio Test

Permalink

https://escholarship.org/uc/item/9r1649tx

Author

Wang, Bushi

Publication Date

2011

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Solving Consistency Problems in Multiple Hypotheses Testing with Consonant Likelihood Ratio Test

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Bushi Wang

March 2011

Dissertation Committee:

Dr. Xinping Cui, Chairperson
Dr. Jun Li
Dr. Tae-Hwy Lee

Copyright by
Bushi Wang
2011

The Dissertation of Bushi Wang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I give special thanks to my advisor, Dr. Xinping Cui, for all of her support, encouragement and guidance throughout my graduate study. I thank Dr. Jason Hsu for his helpful discussions and encouragement. I also thank Dr. Haibing Zhao for his lively discussion and help with Chapter two of my dissertation. I thank Dr. Barry Arnold, Dr. Daniel Jeske, Dr. Subir Ghosh, Dr. Keh-Shin Lii, Dr. Changxuan Mao, Dr. Jun Li, Dr. D.V. Gokhale and Dr. Vanamamalai Seshadri for their excellent guidance in statistics. I thank Paula Lemire, Dr. Linda Penas, Perla Fabelo, Dr. Jason Wilson, Dr. James Flegal for years of help and support. I thank the University of California, Riverside statistics department and my fellow graduate students for five years delighted memory. My thanks also go to Dr. Tae-Hwy Lee, Dr. Thomas Girke and Dr. Tim Close for their work on my final and oral exam committee. My special thanks go to my wife, Dr. Nan Shao, for her constant support and encouragement.

To my wife, Nan Shao and my parents, Dean Wang and Fengyun Zhao.

ABSTRACT OF THE DISSERTATION

Solving Consistency Problems in Multiple Hypotheses Testing with Consonant Likelihood Ratio Test

by

Bushi Wang

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, March 2011
Dr. Xinping Cui, Chairperson

In the past century, several multiple testing procedures have been developed based on the closure principle. Procedures following the closure principle are called closed test and they find individual tests for intersection hypotheses in the family and collates the results to control the familywise error rate. Almost all popular closed tests use the union-intersection method for their intersection hypotheses. However, due to computational issue, likelihood ratio test has not yet received much attention in closed testing. Moreover, not being a union-intersection method, the likelihood ratio test will not satisfy the logical consistency requirement of consonance.

In this dissertation, a general solution (consonance adjustment) to the consistency problem in multiple hypotheses testing is proposed under the framework of closure principle and partitioning principle. Simulation examples are used to show its advantages.

Besides a general framework, also provided in this dissertation is a special form of the consonant test in the context of multiple comparisons with a control. Using likelihood ratio test with the closure principle and a consonance adjustment step, the consonant closed likelihood ratio test is introduced. The rejection region, shortcut and steps in critical constants calculation are discussed. Extensive simulation studies and pre-clinical trial example are provided.

More recently, testing for efficacy in multiple endpoints has emerged as a challenging

statistical problem in clinical trials. Current approaches to this problem are based on closed testing or partition testing, which test the efficacy in certain dose-endpoint combinations and collate the results. Partition testing is in general a more powerful approach since it tests fewer hypotheses to avoid unnecessary power loss. However, all current approaches are still based on various union-intersection tests. In my dissertation, I generalize the decision path principle proposed by Liu and Hsu (2009) to the cases with multiple primary endpoints. Then propose a new partition testing approach based on consonance adjusted likelihood ratio test. The new procedure provides consistent inferences and yet it is still conservative and does not rely on the estimation of endpoint correlations or independence assumption which might be challenged by the regulatory agencies.

Contents

List of Figures **x**

List of Tables **xi**

1 Multiple Testing Procedures: A Review **1**
1.1 Introduction . . . 1
1.1.1 Multiple Hypotheses Testing . . . 1
1.1.2 Closure Principle . . . 2
1.2 Closed Test . . . 3
1.2.1 Holm's step-down test . . . 4
1.2.2 Hochberg's step-up test . . . 5
1.2.3 Hommel's test . . . 6
1.2.4 Step-down Dunnett test . . . 8
1.2.5 Dunnett-Tamhane step-up test . . . 10
1.3 Partitioning Principle . . . 11
1.3.1 Stefansson's partition testing . . . 11
1.3.2 Weak and strong partitioning principle . . . 12

2 General Solutions to Consistency Problems in Multiple Hypothesis Testing **15**
2.1 Coherence, Consonance and Partitioning Principle . . . 17
2.1.1 Coherence and consonance . . . 17
2.1.2 Extension of partitioning principle . . . 20
2.2 Example Illustration and Simulation Study . . . 25
2.2.1 Construct consonant partitioning test . . . 25
2.2.2 Simulation Study . . . 28
2.2.3 Discussion . . . 32

3 Consonant Closed Likelihood Ratio Test Procedures with Application to Dose-Response Study **35**
3.1 Introduction . . . 35
3.2 Step-down Dunnett (SD) Test and Dunnett-Tamhane Step-up (SU) Test as Benchmarks . . . 38
3.3 Consonant closed F test against two-sided alternatives . . . 42
3.3.1 A consonant closed F test in the simple case . . . 42
3.3.2 Compare with SD test and SU test . . . 43
3.3.3 A consonant closed F test comparing k doses with a control . . . 47

3.4 Consonant closed order restricted test against one-sided alternatives 50
3.4.1 Consonant closed order restricted test 50
3.4.2 Compare with SD test and SU test 51
3.5 Dose-response examples 59
3.6 Discussion 61
3.6.1 Consonant closed F test in mixed model with repeated measurements . 61
3.6.2 Computational Issue 64

4 A New Partition Testing Strategy for Multiple Primary and Secondary Endpoints in Regulatory Applications 66
4.1 Introduction to multiple endpoints problem in clinical trials 66
4.2 The Partitioning Decision Path Principle 70
4.2.1 Hierarchical endpoints 70
4.2.2 Alternative primary endpoints 73
4.2.3 Co-primary endpoints 74
4.3 Likelihood Ratio Test and Consonance Adjustment: Fundamentals 76
4.3.1 Likelihood Ratio Test 76
4.3.2 Consonance Adjustment 82
4.4 Likelihood Ratio Test and Consonance Adjustment: Advanced Case 83
4.4.1 Primary-Secondary Endpoints with 3 Doses and a Control 83
4.4.2 Notes for critical constants calculation 84
4.4.3 Simulation study and clinical trial example 87
4.4.4 Co-Primary Endpoints with 2 Doses and a Control 90
4.5 Discussion and Future Research 93

5 Conclusions 96

List of Figures

3.1 The rejection region for SD test. '1' is the region to reject both hypotheses; '2' is the region to reject one but not both hypotheses; '3' is the region to retain both hypotheses. 40
3.2 The rejection region for SU test. '1' is the region to reject both hypotheses; '2' is the region to reject one but not both hypotheses; '3' is the region to retain both hypotheses. 41
3.3 The rejection region for CCF test, SD test (upper half plane only) and SU test (lower half plane only). '1' is the region to reject both hypotheses; '2' is the region to reject one but not both hypotheses; '3' is the region to retain both hypotheses. 45
3.4 The power difference between CCF test and SD test (upper half plane) and SU test (lower half plane) for different (θ_1, θ_2) combinations. A positive value indicates the power advantage of CCF test. The result is based on 10^5 simulations. 46
3.5 The rejection region for CCOR, SD and SU test. '1' is the region to reject both hypotheses; '2' is the region to reject one but not both hypotheses; '3' is the region to retain both hypotheses. 53
3.6 The power comparison among CCOR test and SD test in various (θ_1, θ_2) combinations from 10^5 simulations. 55
3.7 The power comparison among CCOR test and SU test in various (θ_1, θ_2) combinations from 10^5 simulations. 56

4.1 The path partition of the partition testing in the case of one primary endpoint and one secondary endpoint. 71
4.2 The disjointness partition of the partition testing in the case of two doses with one primary endpoint and one secondary endpoint. The labeling of disjointness partition follows that described in (4.3). 72
4.3 The performance comparison when the correlation between primary endpoint and secondary endpoint is 0.3. 89

List of Tables

2.1 The counts of several consistent or inconsistent rejection events for three testing procedures in the first simulation study. No other cases are possible since the procedures compared here are all coherent tests. The computation is based on $100\,000$ simulations with $\theta_1 = -0.4$, $\theta_2 = 0$, $\theta_3 = 0.6$ and $\sigma^2 = 1$. All procedures control the local test at level $\alpha = 0.05$. Sample size $n = 30$ and $n = 10$ (results showed in parentheses) are considered for each population. . . 30

2.2 The counts of several consistent or inconsistent rejection events for three testing procedures in the second simulation study. No other cases are possible since the procedures compared here are all coherent tests. The computation is based on $100\,000$ simulations with $\theta_1 = \theta_2 = 0$, $\theta_3 = 0.8$ and $\sigma^2 = 1$. All procedures control the local test at level $\alpha = 0.05$. Sample size $n = 30$ and $n = 10$ (results showed in parentheses) are considered for each population. 34

3.1 Critical constants of CCF test for selected correlation ρ and degrees of freedom, $\alpha = 0.05$. 49

3.2 Critical constants of CCOR test for selected correlation ρ and degrees of freedom, $\alpha = 0.05$. 52

3.3 Performance of CCOR, SD and SU tests evaluated at selected $\boldsymbol{\theta}$ combinations from $s = 10^6$ simulations, $\alpha = 0.05$. 57

3.4 Performance of CCOR, SD and SU tests in the case of 4 treatments, evaluated at selected $\boldsymbol{\theta}$ combinations from $s = 10^6$ simulations, $\alpha = 0.05$. Column $l/4$ gives the rate of successfully rejecting l false null hypotheses. Column marked 'Avg' gives the average number of correct rejections. 58

3.5 Sample dose-response data from Ruberg (1995) and performance of closed test procedures for selected Δ values. Significant treatment effects are marked with "✓". 59

3.6 FEV1 improvement in asthma trial example 61

4.1 The null hypotheses of interest and null hypotheses actually tested featuring 2 doses and primary-secondary endpoints. A short dash '$-$' indicates no constraint on that parameter. 78

4.2 The natural partition of primary-secondary endpoints with 3 doses and a control. 85

4.3 Scenarios of mean matrices for power comparison 87

4.4 Power comparison for four scenarios with true difference specified in Table 4.3. 88

4.5 Results of a dose-finding study in patients with hypertension. 90

4.6 The natural partition of co-primary endpoints with 2 doses and a control. 91

4.7 The strong partitioning principle applied to decision path partitioning. 94

Chapter 1

Multiple Testing Procedures: A Review

1.1 Introduction

1.1.1 Multiple Hypotheses Testing

Simultaneously testing several hypotheses is a common practice in statistical inferences. For instance, in clinical trials, the efficacy of a treatment drug is evaluated by comparing several dose levels of the drug to a placebo or an active control. In microarray experiments of pharmacogenomics, tens of thousands of genes are often compared simultaneously. If not properly adjusted, the multiplicity issue will inflate the type I error rate of the hypothesis testing. In general, the more hypotheses we test, the larger chance we might erroneously reject some true null hypotheses.

By making the multiplicity adjustment, a multiple testing procedure usually controls the familywise error rate (FWER). FWER is the probability of falsely rejecting any null hypothesis in multiple testing, which can be controlled in the *weak sense* or in the *strong sense*. Weak control of the FWER is defined as controlling the probability of falsely rejecting any hypothesis when all the hypotheses are actually true. Strong control of the FWER is defined as controlling

the probability of falsely rejecting any hypothesis regardless of the number of true hypotheses.

Suppose there is a family of hypotheses $\mathscr{H} = \{H_i \subset \Theta : \ i \in I\}$, where Θ is the parameter space containing all null hypotheses ($\Theta = \cup_{i\in I} H_i$) and I is the finite or infinite index set. We use the same notation H_i for the hypothesis and its parameter space. The hypothesis H_i is true if $\boldsymbol{\theta} \in H_i$. The strong contorl of the FWER requires

$$\sup_{\boldsymbol{\theta}\in\Theta} P_{\boldsymbol{\theta}}\{\text{at least one incorrect assertion}\} \leqslant \alpha.$$

While the weak control requires

$$\sup_{\boldsymbol{\theta}\in\cap_{i\in I} H_i} P_{\boldsymbol{\theta}}\{\text{at least one incorrect assertion}\} \leqslant \alpha,$$

which really controls the FWER in a specific smaller set. Most of the applications in the literature require strong control of the FWER, as it controls the error rate under any true and false hypotheses combinations. In this dissertation, by controlling of FWER, we will always mean strong control unless specified otherwise.

One popular technique to control the FWER strongly is the Bonferroni method. If E_i is the event that the ith assertion is incorrect, $i \in I$, then using the Bonferroni inequality we obtain

$$\text{FWER} = P\big(\bigcup_{i\in I} E_i\big) \leqslant \sum_{i\in I} P(E_i).$$

As a result, if each individual hypothesis H_i is tested at level $\alpha/|I|$ ($|I|$ is the cardinality of the index set I), then $\sum_{i\in I} P(E_i) \leqslant \alpha$ and the FWER is strongly controlled at level α. This idea is also called Bonferroni approximation.

1.1.2 Closure Principle

Closure method developed by Marcus et al. (1976) is a general principle for constructing multiple testing procedures. Multiple testing procedures based on closure principle are called closed test procedures. For testing a finite family of k null hypotheses $\mathscr{H} =$

$\{H_i : i = 1, \ldots, k\}$, a closed test first generate a closure family by including all intersections $H_I = \cap_{i \in I} H_i$, $I \subseteq \{1, \ldots, k\}$. Then it rejects the null hypothesis H_i if all the intersection hypotheses H_J with $J \ni i$ are rejected at individual level α.

The proof is simple, suppose R is the index set of all the true null hypotheses. We will commit a type I error if there exists at least one index $i \in R$ that H_i is rejected. This event, we denote as A, will happen if and only if the intersection hypothesis H_R is rejected. We denote the rejection of H_R as B. Then

$$P(A \cap B) = P(A|B)P(B) \leqslant \alpha,$$

since H_R is rejected by a level α test. Also since $A \cap B = A$, we have $P(A) \leqslant \alpha$.

Different choice of level α test for the intersection hypothesis H_I will result in different closed test procedures. A common practice applied by several closed test procedures is the union-intersection (UI) method proposed by Roy (1953). For any intersection hypothesis $H_I = \cap_{i \in I} H_i$, suppose that tests are available for testing each H_i for $i \in I$. Then the UI test rejects H_I if at least one of the component hypotheses H_i is rejected.

Now the question is how to choose tests for component hypotheses H_i and the level of each test. In the next section, several well-known and powerful closed test procedures that utilize different techniques to test each component hypotheses H_i are reviewed.

1.2 Closed Test

In this section, several closed test procedures are reviewed and compared. They differ mainly in the selection of test for component hypotheses when applying the UI method. Some of the procedures allow step-down shortcuts and some of them have step-up shortcuts depending on the test statistics used in the procedure.

Another difference among these closed tests is that some of them assume normality

in the distributional setting and the others use Bonferroni approximation under the nonparametric framework without specifying the distribution or correlation among the hypotheses. For example, the procedures proposed by Holm, Hochberg and Hommel all use p-value as their test statistics without distributional assumptions. Hochberg and Hommel's procedure assume independence among the p-values while Holm's procedure does not. The step-down Dunnett test and Dunnett-Tamhane step-up test are proposed under the normal assumption with known correlation among test statistics. These procedures will be reviewed in the order as mentioned above.

1.2.1 Holm's step-down test

The first procedure introduced here is Holm's step-down test (Holm, 1979), which use the Bonferroni procedure to test each intersection hypothesis $H_I = \cap_{i \in I} H_i$ for $I \subseteq \{1, \ldots, k\}$. It takes a sequentially rejective form beginning by testing the most significant hypothesis $H_{(k)}$, where the hypotheses are ordered according to their p-values: $p_{(1)} \geqslant p_{(2)} \geqslant \cdots \geqslant p_{(k)}$ with $H_{(k)}$ being the hypothesis with the smallest p-value and (k) being the corresponding index. We call a stepwise procedure starts from the most significant hypothesis "step-down" test.

Holm's step-down test proceeds as follows:

1. Order the p-values p_i: $p_{(1)} \geqslant p_{(2)} \geqslant \cdots \geqslant p_{(k)}$. Let $H_{(1)}, \ldots, H_{(k)}$ be the corresponding null hypotheses.

2. Reject $H_{(k)}$ if $p_{(k)} < \alpha/k$ and go to the next step. Otherwise retain all hypotheses and stop.

3. Reject $H_{(k-1)}$ if $p_{(k-1)} < \alpha/(k-1)$ and go to the next step. Otherwise retain $H_{(k-1)}, \ldots, H_{(1)}$ and stop, etc.

Obviously, hypothesis $H_{(i)}$ is rejected if and only if $p_{(j)} < \alpha/j$ for all $j = k, k -$

$1, \ldots, i$. Less obviously, this condition indicates that for any index set $I \ni (i)$, $\min_{l \in I} p_l < \alpha/|I|$. This is because for any index set I, if $(i) \in I$, $\min_{l \in I} p_l \leqslant p_{(i)}$. Suppose $\min_{l \in I} p_l = p_{(t)}$ for some specific $t \in \{k, k-1, \ldots, i\}$. Since for all $j = k, k-1, \ldots, i$, $p_{(j)} < \alpha/j$, we have $\min_{l \in I} p_l = p_{(t)} < \alpha/t$. Also notice that if $\min_{l \in I} p_l = p_{(t)}$, then the index $(t+1), \ldots, (k)$ are not included in the index set I and the cardinality $|I| \leqslant t$, thus $\min_{l \in I} p_l = p_{(t)} < \alpha/t \leqslant \alpha/|I|$.

This result shows that hypothesis $H_{(i)}$ is rejected if and only if for any index set $I \ni (i)$, $\min_{l \in I} p_l < \alpha/|I|$. The step-down procedure origins from a closed test procedure which reject each intersection hypothesis H_I if $\min_{l \in I} p_l < \alpha/|I|$, which is a Bonferroni procedure. It can be also considered as a UI method since H_I is rejected if at least one of the component hypotheses has a p-value less than $\alpha/|I|$.

1.2.2 Hochberg's step-up test

On the contrary, Hochberg's step-up test (Hochberg, 1988) starts from the least significant hypothesis $H_{(1)}$ with the largest p-value $p_{(1)}$. The procedure works as follows:

1. Order the p-values p_i: $p_{(1)} \geqslant p_{(2)} \geqslant \cdots \geqslant p_{(k)}$. Let $H_{(1)}, \ldots, H_{(k)}$ be the corresponding null hypotheses.

2. Retain $H_{(1)}$ if $p_{(1)} \geqslant \alpha$ and go to the next step. Otherwise reject all hypotheses and stop.

3. Retain $H_{(2)}$ if $p_{(2)} \geqslant \alpha/2$ and go to the next step. Otherwise reject $H_{(2)}, \ldots, H_{(k)}$ and stop, etc.

In general, hypothesis $H_{(i)}$ is retained if and only if $p_{(j)} \geqslant \alpha/j$ for $j = 1, 2, \ldots, i$. Equivalently, hypothesis $H_{(i)}$ will be rejected if and only if for some $j \in \{1, 2, \ldots, i\}$, $p_{(j)} < \alpha/j$. Less obviously, this is also equivalent to reject $H_{(i)}$ if for any index set $I \ni (i)$, the following condition is met:

$$p_{(1)^I} < \alpha \text{ or } p_{(2)^I} < \alpha/2 \text{ or } \ldots \text{ or } p_{(|I|)^I} < \alpha/|I|. \tag{1.1}$$

The p-values $p_{(1)^I} \geqslant p_{(2)^I} \geqslant \ldots \geqslant p_{(|I|)^I}$ are ordered within the index set I. This is true because there exist at least one $j \in \{1, 2, \ldots, i\}$ such that $p_{(j)} < \alpha/j$. This index (j) has a new rank in the subset $I \subset \{1, 2, \ldots, k\}$, and the new rank t can not exceed j since there are fewer indices in the set I than $\{1, 2, \ldots, k\}$. As $\alpha/j \leqslant \alpha/t$, $p_{(t)^I} = p_{(j)} < \alpha/j \leqslant \alpha/t$ and (1.1) is satisfied.

This concludes that Hochberg's step-up test origins from a closed test procedure which test each intersection hypothesis H_I if at least one of the inequalities in (1.1) is satisfied. This procedure can be also viewed as a UI method since H_I is rejected if at least one of the component hypotheses is rejected by comparing its p-value to some specific constants.

1.2.3 Hommel's test

Another variation of the closed test is proposed by Hommel (Hommel, 1988), which does not allow a stepwise shortcut. The procedure also orders the p-values as $p_{(1)} \geqslant p_{(2)} \geqslant \cdots \geqslant p_{(k)}$ and finds the largest value of m ($1 \leqslant m \leqslant k$) such that

$$p_{(j)} > (m - j + 1)\alpha/m, \quad j = 1, 2, \ldots, m.$$

If there is no such m then it rejects all hypotheses; else it rejects those H_i with $p_i < \alpha/m$. Suppose such m exists, then the following inequalities are satisfied:

$$\begin{aligned} p_{(1)} &> \alpha \\ p_{(2)} &> \frac{m-1}{m}\alpha > \frac{m-2}{m-1}\alpha > \cdots > \frac{1}{2}\alpha \\ &\vdots \\ p_{(m)} &> \frac{1}{m}\alpha \\ p_{(m+1)} &\leqslant \frac{1}{m+1}\alpha < \frac{1}{m}\alpha \\ &\vdots \\ p_{(k)} &\leqslant \frac{1}{k}\alpha < \cdots < \frac{1}{m}\alpha. \end{aligned} \tag{1.2}$$

The hypotheses $H_{(1)}, \ldots, H_{(m)}$ are to be retained and the rest will be rejected. It has been pointed out by Huang and Hsu (2007) that Hommel's test is essentially a closed test which rejects each intersection hypothesis H_I if

$$p_{(i)^I} < (|I| - i + 1)\alpha/|I|, \text{ for some } i, \tag{1.3}$$

with $p_{(i)^I}$ being the ordered p-value among the null hypotheses involved in H_I. It is also a UI method which rejects the intersection hypothesis H_I if at least one of the component hypotheses is rejected by comparing its p-value to a pre-specified constant.

The above statement can be verified if we consider two scenarios separately, $i \leqslant m$ and $m < i \leqslant k$. In the first scenario, consider hypothesis $H_{(i)}$ with $i \leqslant m$. Simply pick any index set $I \ni (i)$ and $|I| \leqslant m$, which only contains indices in $\{(1), \ldots, (m)\}$. Then all the p-values $p_{(j)^I} > (|I| - j + 1)\alpha/|I|$ according to (1.2) and H_I will be retained according to (1.3). As a result, the hypothesis $H_{(i)}$ will also be retained.

In the second scenario, consider hypothesis $H_{(i)}$ with $m < i \leqslant k$. For any index set $I \ni (i)$, there are two possibilities, $|I| \leqslant i$ or $|I| > i$. In the case $|I| \leqslant i$, suppose the p-value

of $H_{(i)}$ has a new order j in the set I. Then $p_{(i)} = p_{(j)^I} < \alpha/i$ according to (1.2). Due to the fact that $j \leqslant |I| \leqslant i$, we have $p_{(i)} = p_{(j)^I} < \alpha/i \leqslant \frac{|I|-j+1}{|I|}\alpha$. According to (1.3), the intersection hypothesis H_I should be rejected. In the case $|I| > i > m$, the smallest p-value in the set is $p_{(l)} = p_{(|I|)^I}$ where $l \geqslant |I|$. Thus according to (1.2), $p_{(l)} = p_{(|I|)^I} < \alpha/l \leqslant \alpha/|I|$ and the hypothesis H_I should be rejected based on (1.3). Therefore $H_{(i)}$ should be rejected in the second scenario.

Now we see the procedure based on (1.3) gets the same result to retain hypotheses $H_{(1)}, \ldots, H_{(m)}$ and reject hypotheses $H_{(m+1)}, \ldots, H_{(k)}$. Thus Hommel's test is a closed test using UI test for each intersection hypothesis.

1.2.4 Step-down Dunnett test

The rest two subsections introduce two parametric procedures assuming normality for inter-treatment comparisons. Consider the following distributional setting. For $i = 1, 2, \ldots, k$, let the θ_i be the parameters of interest in a general linear model and let $\hat{\theta}_i$ be their least squares estimators. We assume that $\hat{\theta}_i$ are jointly normally distributed with means θ_i, variance $\tau_i^2\sigma^2$ and correlation coefficient Corr$(\hat{\theta}_i, \hat{\theta}_j) = \rho_{ij}$. The τ_i^2 and ρ_{ij} are known design-dependent constants and σ^2 is an unknown scalar. Simultaneous inferences are desired on the θ_i's. We also assume that S^2 is an unbiased mean square error estimator of σ^2 with ν degrees of freedom and $\nu S^2/\sigma^2 \sim \chi^2_\nu$ independent of the $\hat{\theta}_i$'s.

A very common application in pharmaceutical statistics is to compare several treatments with a control. Label the treatments by $0, 1, 2, \ldots, k$ where 0 is the control. Denote by $\bar{Y}_i$ the sample mean based on n_i observations, then $\bar{Y}_i$ are assumed to be independent $N(\mu_i, \sigma^2/n_i)$ random variables. The parameters of interest are $\theta_i = \mu_i - \mu_0$ for $i = 1, 2, \ldots, k$ with estimator

being $\bar{Y}_i - \bar{Y}_0$. The unbiased mean squared error estimator of σ^2 is

$$S^2 = \frac{\sum_{i=0}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{\sum_{i=0}^{k} (n_i - 1)} \tag{1.4}$$

with $\nu = \sum_{i=0}^{k} (n_i - 1)$ degrees of freedom and $\nu S^2/\sigma^2 \sim \chi^2_\nu$ independent of $\bar{Y}_i$. It is easy to show that

$$\rho_{ij} = \text{Corr}(\bar{Y}_i - \bar{Y}_0, \bar{Y}_j - \bar{Y}_0) = \lambda_i \lambda_j, \quad \text{where } \lambda_i = \sqrt{\frac{n_i}{n_i + n_0}}, \quad i \neq j \in \{1, \ldots, k\}.$$

And $\tau_i^2 = 1/n_i + 1/n_0$ for $i = 1, 2, \ldots, k$. To test the hypotheses $H_i : \theta_i = 0$ vs. $A_i : \theta_i > 0$, the test statistics are

$$T_i = \frac{\bar{Y}_i - \bar{Y}_0}{\tau_i S}, \quad i = 1, \ldots, k. \tag{1.5}$$

For each intersection hypothesis $H_I = \cap_{i \in I} H_i$, Marcus et al. (1976) applied the union-intersection method and test H_I with Dunnett's single step test, which rejects H_I if at least one of H_i with $i \in I$ has a large test statistic.

$$\max_{i \in I} T_i > t_I, \quad i \in I \tag{1.6}$$

and t_I is the solution in t of the equation:

$$\int_0^\infty \int_{-\infty}^{\infty} \prod_{i \in I} \Phi \left(\frac{\lambda_i z + tu}{\sqrt{1 - \lambda_i^2}} \right) d\Phi(z) dF_\nu(u) = 1 - \alpha, \tag{1.7}$$

where $\Phi(\cdot)$ is the standard normal cdf and $F_\nu(\cdot)$ is the cdf of a $\sqrt{\chi^2_\nu/\nu}$ random variable. Note that t_I depends on the index set I as well as α, ν and $\{\rho_{ij}\}$.

Notice that this procedure is similar to Holm's step-down test except using a maximum t statistic instead of minimum p-values. It also takes a particularly simple step-down form when $n_1 = n_2 = \ldots = n_k = n$, in which case $\lambda_i \equiv \lambda = \sqrt{n/(n + n_0)}$ and $\rho_{ij} \equiv \rho$ for all $i \neq j \in \{1, \ldots, k\}$. The critical constants t_I are simplified to $t_{|I|}$ which depends on I only through its cardinality $|I|$. The steps of the step-down Dunnett (SD) test are as follows:

1. Order the test statistics T_i: $T_{(1)} \leqslant \ldots \leqslant T_{(k)}$. Let $H_{(1)}, \ldots, H_{(k)}$ be the corresponding null hypotheses.

2. Reject $H_{(k)}$ if $T_{(k)} > t_k$ and go to next step. Otherwise retain all hypotheses and stop.

3. Reject $H_{(k-1)}$ if $T_{(k-1)} > t_{k-1}$ and go to next step. Otherwise retain $H_{(k-1)}, \ldots, H_{(1)}$ and stop, etc.

Obviously t_1 equals to the upper α quantile of the t distribution with ν degrees of freedom and $t_1 < t_2 < \ldots < t_k$ since the integrand in (1.7) is decreasing in $|I|$. An extension of this step-down test for the case of unequal sample size when comparing treatments with a control is given by Dunnett and Tamhane (1991). Holm's step-down test can be viewed as Bonferroni approximation to SD test.

1.2.5 Dunnett-Tamhane step-up test

Dunnett and Tamhane (1992) proposed a step-up (SU) multiple testing procedure analogous to Hochberg's step-up test for the case of equal sample size. It is also a closed test procedure in essence (Liu, 1996). Instead of using the maximum t statistics in (1.6), their procedure tests the intersection hypothesis $H_I = \cap_{i \in I} H_i$ using all ordered t statistics and a common set of critical values. Suppose $T_{(1)^I} \leqslant \ldots \leqslant T_{(|I|)^I}$ are the ordered t statistics of index set I, the procedure rejects $H_I = \cap_{i \in I} H_i$ if

$$T_{(1)^I} > c_1 \text{ or } T_{(2)^I} > c_2 \text{ or } \ldots \text{ or } T_{(|I|)^I} > c_{|I|}, \tag{1.8}$$

where the critical constant $c_1 \leqslant \ldots \leqslant c_k$ are solutions to the recursive equations:

$$P\{T_{(1)} \leqslant c_1, \ldots, T_{(m)} \leqslant c_m\} = 1 - \alpha, \quad m = 1, \ldots, k \tag{1.9}$$

and $T_{(1)} \leqslant \ldots \leqslant T_{(m)}$ are the order statistics of $T_1, \ldots, T_m$.

This procedure can be simplified to the following step-up version:

1. Order the test statistics T_i: $T_{(1)} \leqslant \ldots \leqslant T_{(k)}$. Let $H_{(1)}, \ldots, H_{(k)}$ be the corresponding null hypotheses. Compute critical constants $c_1 \leqslant \ldots \leqslant c_k$ as indicated above.

2. Retain $H_{(1)}$ if $T_{(1)} \leqslant c_1$ and go to next step. Otherwise reject all hypotheses and stop.

3. Retain $H_{(2)}$ if $T_{(2)} \leqslant c_2$ and go to next step. Otherwise reject $H_{(2)}, \ldots, H_{(k)}$ and stop, etc.

Dunnett and Tamhane (1995) also give an extension of this procedure for the case of unequal sample size. However, there is no analytical proof that it controls FWER strongly, which is the major drawback of SU test.

1.3 Partitioning Principle

Partitioning principle was first proposed by Stefansson et al. (1988). With all else being equal, partition testing is as least as powerful as closed testing. The advantage of partitioning principle is that the parameter space is partitioned into disjoint pieces and no multiplicity adjustment is needed thereafter.

1.3.1 Stefansson's partition testing

Consider testing a finite family of k hypotheses $\mathscr{H} = \{H_i : \ i = 1, \ldots, k\}$. To control the FWER strongly, the partitioning principle first partitions the entire parameter space $\cup_{i=1}^{k} H_i$ into disjoint $H_I^\star$ for $I \subseteq \{1, 2, \ldots, k\}$ as follows. For each $I \subseteq \{1, 2, \ldots, k\}$, let

$$H_I^\star = \big(\cap_{i \in I} H_i\big) \bigcap \big(\cap_{j \notin I} H_j^c\big),$$

with H_j^c being the complement of H_j. Then $\{H_I^\star : I\}$ including $H_\varnothing^\star = \cap_{i=1}^{k} H_i^c$ partition the parameter space. In fact, each partition $H_I^\star$ can be interpreted as the parameter space that exactly H_i for $i \in I$ are true and the rest hypotheses are false.

Since all $H_I^\star$ are disjoint, at most one of them contains the true parameter(s). Test each one of the partitions at level α and the probability of rejecting at least one true null hypothesis is at most α without any multiplicity adjustment.

The partition testing reject any original null hypothesis H_i if all the partition $H_I^\star$ with $I \ni i$ are rejected.

Partition testing procedures have been applied in the literature in the areas such as bioequivalence test (Berger and Hsu, 1996), dose-response studies (Hsu and Berger, 1999) and genetic linkage analysis (Lin et al., 2001).

1.3.2 Weak and strong partitioning principle

Finner and Strassburger (2002) further refined the partitioning principle and proposed the weak and strong partitioning principle by putting restrictions on where to look for the least favorable configurations.

Theorem 1 (Weak Partitioning Principle, WPP) *Let $\mathscr{H} = \{H_i : \ i \in K\}$ be a finite family of k hypotheses where $K = \{1, 2, \ldots, k\}$. Let $\mathscr{H}^\star = \{H_I^\star : \ I \subseteq K\}$ denote the partition of the parameter space $\cup_{i=1}^k H_i$. For each $I \subseteq K$ choose test φ_I for testing H_I such that*

$$\sup_{\boldsymbol{\theta} \in H_I^\star} P_{\boldsymbol{\theta}}(\varphi_I = 1) \leqslant \alpha$$

Define $\phi = (\phi_i : i \in K)$ by

$$\phi_i = \min_{I \subseteq K : I \ni i} \varphi_I, \quad \forall i \in K$$

Then ϕ controls the FWER strongly.

In contrast to the closed testing procedure, the level α test φ_I for testing H_I are chosen such that the type I error are only controlled over $H_I^\star \subset H_I$. That is, the procedure looks for the least favorable configuration over $H_I^\star$ instead of H_I as one would do applying the

closure principle. As a result this may lead to a test with a smaller acceptance region, hence a more powerful test for H_I. The WPP is at least as powerful as the closure principle. However, whether the WPP indeed yields a more powerful test procedure than the closure principle heavily depends on the structure of $\mathscr{H}$ and the structure of the acceptance regions.

A more advanced and sometimes more powerful method is the strong partitioning principle.

Theorem 2 (Strong Partitioning Principle, SPP) *Let $\mathscr{H} = \{H_i : i \in K\}$ be a finite family of k hypotheses where $K = \{1, 2, \ldots, k\}$. Let $\mathscr{H}^\star = \{H_I^\star : I \subseteq K\}$ denote the partition of the parameter space $\cup_{i=1}^k H_i$. For each $I \subseteq K$ choose test φ_I for testing H_I such that*

$$\forall I \subseteq K : \quad \varphi_I = \min_{J \subseteq K : J \subseteq I} \varphi_J \tag{1.10}$$

and

$$\forall I \subseteq K : \quad \sup_{\boldsymbol{\theta} \in H_I^\star} P_{\boldsymbol{\theta}}(\varphi_I = 1) \leqslant \alpha$$

Then $\{\varphi_i : i \in K\}$ is a multiple testing procedure strongly controls the FWER at α.

The difference between SPP and WPP is the additional requirement (1.10) for the local level α tests φ_I for H_I which ensures the consistency condition discussed in the next chapter. To sum up, weather the SPP leads to an improved multiple decision procedure compared with WPP also depends on the hypotheses of interest.

After reviewing all the popular multiple testing procedures, we see that a very important decision for both closed test and partitioning principle is the selection of individual level α test for intersection hypothesis H_I ($H_I^\star$ in partitioning principle). All the reviewed closed testing procedures choose to use some form of an union-intersection test. However, we have noticed in the introduction to partitioning principle that we are not confined to use union-intersection

test. A very common non-union-intersection test is the likelihood ratio test. We will see in the following chapters that in lieu of the union-intersection test, we need to be careful with some consistency issues among multiple inferences.

In Chapter 2, a general solution to construct logically consistent multiple testing procedures under the partitioning principle framework is proposed. As we know, partitioning principle is a more general method than closure principle and sometimes may yield more powerful multiple testing procedures, the consistent multiple testing procedure can also work under the closed testing framework. In Chapter 3, a specific consistent procedure using the likelihood ratio test is proposed under the closed testing framework for multiple comparisons with a control problem. In Chapter 4, we extend the consistent testing procedure to the problem of dose-response clinical trials with multiple endpoints. The method is implemented with the partition testing framework.

Chapter 2

General Solutions to Consistency Problems in Multiple Hypothesis Testing

Consider simultaneously testing several different but possibly related null hypotheses, $H_i, i \in I$. In general, it is relatively easy to construct a powerful test at level α for each individual hypothesis H_i. However, combining these individual tests into a multiple testing procedure (MTP) with desired familywise error rate can be challenging as the MTP has to be powerful and make no contradictory conclusions.

A variety of multiple testing procedures have been proposed in the past fifty years. Among them, closure principle has been first shown as a powerful tool to construct multiple level-α tests. It has been shown (see for example Finner and Strassburger (2002), Theorem 2.2) that coherence is a minimal requirement for multiple tests to avoid contradictory conclusions. *An MTP is coherent if not rejecting hypothesis H_i will result in not rejecting any H_j which is implied by H_i* (H_j is implied by H_i means $H_i \subseteq H_j$). Any MTPs based on closure principle are

coherent. Partitioning principle provides another powerful tool for constructing MTPs. MTPs constructed under partitioning principle are also coherent.

It should be emphasized that ensuring an MTP to be *consonant* is also important from both interpretive and mathematical statistics viewpoint (Bittman et al., 2009). However, consonance has not yet received as much attention and discussion as coherence. *A multiple testing procedure is consonant if rejecting hypothesis H_i will result in rejecting at least one of H_j which is implied by H_i.* As pointed out by Bittman et al. (2009), a nonconsonant or dissonant testing procedure can leave the statistician in a difficult situation when explaining the results of a study. For example, in comparing two treatment effects with a control, it is hard to make interpretation if the ANOVA shows significant treatment effect but none of the contrast test is significant to declare pairwise difference. Lack of consonance makes interpretation awkward. Bittman et al. (2009) also argued from a mathematical statistics viewpoint that dissonance is undesirable in that it results in decreased ability to reject false null hypotheses.

Following the discussion in the previous chapter, if union-intersection tests are applied to the intersection hypotheses in closed testing procedures or partition testing procedures, the consonance requirement will be satisfied. However, if non-union-intersection methods are used, the likelihood ratio test for example, will result in dissonant testing procedures.

Sonnemann (2008) [the translated version of Sonnemann (1982)] presented a simple method, which will not reject H_i if none of its supersets are rejected, to adjust dissonant procedure to be consonant. Obviously, this method is not powerful. Bittman et al. (2009) proposed a maximin test $\{\varphi_i, i \in I\}$, which is consonant, for testing $\{H_i, i \in I\}$. However, they did not provide a general framework for the construction of MTPs that are both consonant and coherent. Moreover, issues of constructing strongly consonant (Sonnemann, 2008) and strongly coherent (Finner, 1994) MTPs were not considered in their work.

In this chapter, building upon the partitioning principle in Finner and Strassburger

(2002), we propose a general framework, which consists of several adjustment methods, for constructing consonant (strongly consonant) and strongly coherent MTPs. We show that the new methods are more powerful than the method proposed in Sonnemann (2008).

In section 2.1, the concepts of coherence, strong coherence, consonance and strong consonance are reviewed first, followed by the methods to construct consonant tests and strongly consonant tests through partitioning principle. In section 2.2, the closure principle, partitioning principle and the proposed consonant partitioning principle are illustrated and compared in several examples with some discussion.

2.1 Coherence, Consonance and Partitioning Principle

2.1.1 Coherence and consonance

Suppose $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is a family of probability measures defined on a common sample space $(\mathcal{X}, \mathscr{B})$. Let $\mathscr{H} = \{H_i : i \in I\}$ be a family of (null) hypotheses with $\varnothing \neq H_i \subset \Theta$, where I is any index set and Θ is a parameter space. The alternative hypotheses are defined by $K_i = \Theta \backslash H_i$. The index set $I(\theta) = \{i \in I : \theta \in H_i\}$ denotes the set of all indices of true null hypotheses if θ is the true parameter. The family $\mathscr{H}$ is said to be $\cap$-closed if for all $\varnothing \neq J \subseteq I$ either $\bigcap_{j\in J} H_j \in \mathscr{H}$ or $\bigcap_{j\in J} H_j = \varnothing$.

A (non-randomized) multiple test $\varphi = (\varphi_i : \mathcal{X} \to \{0, 1\}, i \in I)$ for $\mathscr{H}$ is said to be a multiple level α test if

$$\forall \theta \in \Theta : \; P_\theta\Big(\bigcup_{i\in I(\theta)} \{\varphi_i = 1\} \Big) \leqslant \alpha,$$

and is denoted by $\varphi \in \Phi_\alpha(\mathscr{H})$. A multiple test $\varphi \in \Phi_\alpha(\mathscr{H})$ is said to be

1. *coherent*, if

$$\forall i, j \in I : \; [H_i \subset H_j \Rightarrow \varphi_i \geqslant \varphi_j];$$

2. *consonant,* if

$$\forall i \in I, \forall x \in \mathcal{X} : [\varphi_i(x) = 1 \text{ and } \exists H_j \supset H_i \Rightarrow \max_{r:H_r \supset H_i} \varphi_r(x) = 1];$$

3. *strongly consonant,* if

$$\forall i \in I, \forall x \in \mathcal{X}, \forall J \in \mathcal{J}(i) = \{\varnothing \neq L \subset I \backslash \{i\} : \bigcap_{j \in L} H_j = H_i\} :$$

$$[\varphi_i(x) = 1 \Rightarrow \max_{j \in J} \varphi_j(x) = 1].$$

Gabriel (1969) provided the first two definitions and the strong consonance is defined by Sonnemann (2008). To show the difference between consonance and strong consonance, we borrow the example from Sonnemann (2008). Consider the problem of comparing three means in a pairwise and global manner in an analysis of variance:

$$H_1 : \theta_1 = \theta_2, \quad H_2 : \theta_1 = \theta_3, \quad H_3 : \theta_2 = \theta_3, \quad H_4 : \theta_1 = \theta_2 = \theta_3.$$

Multiple test $\varphi = (\varphi_1 = 1, \varphi_2 = 0, \varphi_3 = 0, \varphi_4 = 1)$ is consonant since one of H_1, H_2, H_3 is rejected along with H_4. However it is not strongly consonant since $H_2 \cap H_3 = H_4$ but neither H_2 nor H_3 is rejected while $\varphi_4 = 1$. Strong consonance actually requires at least two rejections among H_1, H_2, H_3, thus represents a stronger constraint over a multiple test. To incorporate this constraint in the construction of a more powerful multiple test, we need to distinguish between the following two index sets:

$$\begin{aligned} C(i) &= \{j \in I : H_j \supset H_i, \nexists \text{(do not exist) } H_l \in \mathscr{H} \text{ s.t. } H_i \subset H_l \subset H_j\}, \\ E(i) &= \{j \in I : H_j \supset H_i, \text{ and } H_j \text{ is an elementary hypothesis}\}, \end{aligned} \tag{2.1}$$

where according to Sonnemann (2008), a hypothesis $H_i \in \mathscr{H}$ is called *maximal* or *elementary* hypothesis if do not exit $H_j \in \mathscr{H}$ with $H_j \supset H_i$. Thus $C(i)$ is the index set of all closest supersets of H_i. $E(i)$ is the index set of all elementary supersets of H_i.

The following fundamental lemma shows how to construct a consonant or strongly consonant test from any multiple level α test.

Lemma 3 *Let $\varphi \in \Phi_\alpha(\mathscr{H})$ and $\phi = \{\phi_i : i \in I\}$, $\overline{\phi} = \{\overline{\phi}_i : i \in I\}$ with*

$$\phi_i = \begin{cases} \varphi_i, & \text{if } C(i) = \varnothing; \\ \min(\max\limits_{j \in C(i)} \phi_j,\ \varphi_i), & \text{otherwise.} \end{cases} \qquad \forall i \in I.$$

$$\overline{\phi}_i = \begin{cases} \varphi_i, & \text{if } C(i) = \varnothing\ (\mathcal{J}(i) = \varnothing); \\ \min(\min\limits_{J \in \mathcal{J}(i)} \max\limits_{j \in J} \varphi_j,\ \varphi_i), & \text{otherwise.} \end{cases} \qquad \forall i \in I.$$

where $\mathcal{J}(i) = \{\varnothing \neq L \subset I \backslash \{i\} : \bigcap_{j \in L} H_j = H_i\}$. Then $\phi \in \Phi_\alpha(\mathscr{H})$ and is consonant, $\overline{\phi} \in \Phi_\alpha(\mathscr{H})$ and is strongly consonant.

Proof of Lemma 3. Note that $\phi \leqslant \varphi$, $\overline{\phi} \leqslant \varphi$ and $\varphi \in \Phi_\alpha(\mathscr{H})$. So $\phi, \overline{\phi} \in \Phi_\alpha(\mathscr{H})$ follows easily.

Since $\forall i \in I$ with $C(i) \neq \varnothing$ and $\phi_i \leqslant \max_{j \in C(i)} \phi_j \leqslant \max_{r: H_r \supset H_i} \phi_r$, ϕ is consonant.

To show $\overline{\phi}$ is strongly consonant, it is equivalent to show that $\forall i \in I, \forall J \in \mathcal{J}(i) \neq \varnothing$, if $\overline{\phi}_i = 1$, then $\max_{j \in J} \overline{\phi}_j = 1$.

For $\mathcal{J}(i) \neq \varnothing$, $\overline{\phi}_i = \min(\min_{J \in \mathcal{J}(i)} \max_{j \in J} \varphi_j,\ \varphi_i) = 1$. Next we show $\max_{j \in J} \overline{\phi}_j = 1$, $\forall J \in \mathcal{J}(i)$. If there exist $J_0 \in \mathcal{J}(i)$ such that $\max_{j \in J_0} \overline{\phi}_j = 0$, then

$$\max_{j \in J_0} \min(\min_{K \in \mathcal{J}(j)} \max_{k \in K} \varphi_k,\ \varphi_j) = 0.$$

We split J_0 into two disjoint parts: $J_0 = J_{01} \cup J_{02}$, where $\varphi_j = 1$ if $j \in J_{01}$ and $\varphi_j = 0$ if $j \in J_{02}$. Since

$$\max_{j \in J_{01}} \min(\min_{K \in \mathcal{J}(j)} \max_{k \in K} \varphi_k,\ \varphi_j) \leqslant \max_{j \in J_0} \min(\min_{K \in \mathcal{J}(j)} \max_{k \in K} \varphi_k,\ \varphi_j) = 0,$$

there exist $K_j \in \mathcal{J}(j)$ such that $\max_{k \in K_j} \varphi_k = 0$ for any $j \in J_{01}$. Note that $\bigcap_{j \in J_0} H_j = H_i$ and $\bigcap_{k \in K_j} H_k = H_j$. Then $H_i = \{\bigcap_{j \in J_{01}} \bigcap_{k \in K_j} H_k\} \bigcap \{\bigcap_{j \in J_{02}} H_j\}$. Thus if we denote $\{\bigcup_{j \in J_{01}} K_j\} \bigcup J_{02}$ by J_1, then $J_1 \in \mathcal{J}(i)$. Obviously, $\max_{j \in J_1} \varphi_j = 0$ following the prop-

erty of K_j and J_{02}. This contradicts with $\min(\min_{J\in\mathcal{J}(i)} \max_{j\in J} \varphi_j,\ \varphi_i) = 1$ and we have $\max_{j\in J} \overline{\phi}_j = 1$. ■

Remark. For elementary hypotheses H_i with $E(i) = \varnothing$ ($C(i) = \varnothing$), $\varphi_i = \phi_i = \overline{\phi}_i$, thus there is no cost to adopt consonant test as to the elementary hypotheses (usually the hypotheses of most interest). As for the whole multiple testing procedure, the consonant test retains those null hypotheses, of which rejection can cause inconsistency. As a result, there exists some expendable cost to the consonance. We will see in the following theorems that by excluding those inconsistent rejections, we can reject more false null hypotheses while control the familywise error rate at the same level.

2.1.2 Extension of partitioning principle

Finner and Strassburger (2002) showed that the application of partitioning principle can sometimes lead to more powerful multiple testing procedures that are strongly coherent. A *strongly coherent* multiple test φ is defined in Finner (1994) as

$$\forall x \in \mathcal{X} \text{ and } i \in I:\ [H_i \subseteq \bigcup_{j:\varphi_j(x)=1} H_j \Rightarrow \varphi_i(x) = 1].$$

The following definition of *natural partition* given by Finner and Strassburger (2002) will be needed in constructing new consonant test. Suppose that $\mathscr{H} = \{H_i : i \in I\}$ is $\cap-$closed. Let $\Theta_i = H_i \cap (\cup_{j:H_j\subset H_i} H_j)^c$ for $i \in I$ and let $J_p = \{i \in I : \Theta_i \neq \varnothing\}$. Then $\Theta(J_p) = \{\Theta_i : i \in J_p\}$ is called the natural partition of $\cup_{i\in I} H_i$.

Finner and Strassburger (2002) used the coherence condition along with the natural partition of $\cup_{i\in I} H_i$ to obtain a strongly coherent multiple test. The procedures are named weak and strong partitioning principle according to the calculation of critical values. Following the similar idea, we use consonance condition along with partitioning principle to construct a multiple test that is not only strongly coherent, but also (strongly) consonant, as outlined in the

following three theorems.

Theorem 4 (General Consonant Partitioning Principle) *Suppose that $\mathscr{H} = \{H_i : i \in I\}$ is $\cap$-closed. Let $\Theta(J_p) = \{\Theta_i : i \in J_p\}$ be the natural partition of $\cup_{i\in I} H_i$. For each $i \in J_p$ choose test φ_i for testing H_i versus K_i such that*

$$\sup_{\theta\in\Theta_i} P_\theta(\varphi_i = 1) \leqslant \alpha.$$

Set

$$\overline{\varphi}_i = \min_{j\in J_p:\Theta_j\subseteq H_i} \varphi_j \quad \forall i \in I,$$

$$\phi_i = \begin{cases} \overline{\varphi}_i, & \text{if } C(i) = \varnothing; \\ \min(\max\limits_{l\in C(i)} \phi_l,\ \overline{\varphi}_i), & \text{otherwise.} \end{cases} \qquad \forall i \in I.$$

Then $\phi \in \Phi_\alpha(\mathscr{H})$ and it is strongly coherent and consonant.

Proof of Theorem 4. Most part of Theorem 2.3 follows directly from Lemma 2.1. We only need to show that ϕ is strongly coherent. If we can show $\phi_i = \min_{j\in J_p:\Theta_j\subseteq H_i} \phi_j$, then ϕ is strongly coherent by Theorem 3.2 in Finner and Strassburger (2002). Obviously, we only need to show $\phi_i \leqslant \min_{j\in J_p:\Theta_j\subseteq H_i} \phi_j$.

By the closure of $\mathscr{H}$ and the definition of natural partition, $\varnothing \neq \Theta_j \subseteq H_i$ implies $H_j \subseteq H_i$. When $H_j \subset H_i$, we can always find a series of closest superset hypotheses $H_j \subset H_{j1} \subset \ldots \subset H_{jq} \subset H_i$ connecting H_j and H_i. Thus without loss of generality, we can make the assumption that H_i is the closest superset of H_j , or in other words, $i \in C(j)$. Suppose $\phi_i = 1$, then by the definition of ϕ_i, we have $1 = \overline{\varphi}_i = \min_{l\in J_p:\Theta_l\subseteq H_i} \varphi_l \leqslant \min_{l\in J_p:\Theta_l\subseteq H_j\subset H_i} \varphi_l = \overline{\varphi}_j$. Also since $i \in C(j)$, $\max_{l\in C(j)} \phi_l \geqslant \phi_i = 1$. Combine the arguments together with the definition of $\phi_j = \min(\max_{l\in C(j)} \phi_l, \overline{\varphi}_j)$, we have $\phi_j = 1$ for $j \in J_p : \Theta_j \subseteq H_i$. And ϕ is strongly coherent. ■

Theorem 4 can be viewed as applying the consonance adjustment to Finner and Strassburger (2002)'s weak partitioning principle. The critical values of local tests remain the same. This procedure is more powerful than the multiple consonant test in Sonnemann (2008) when the partitioning principle leads to more powerful local tests.

In analogy to the strong partitioning principle, we put consonance constraint instead of coherence constraint on the local test φ to find the test boundaries as follows:

Theorem 5 (Consonant Partitioning Principle) *Suppose that $\mathscr{H} = \{H_i : i \in I\}$ is $\cap$-closed. Let $\Theta(J_p) = \{\Theta_i : i \in J_p\}$ be the natural partition of $\cup_{i\in I} H_i$. For each $i \in J_p$ choose test φ_i for testing H_i (or Θ_i) versus K_i such that*

$$\varphi_i = \min\left(\max_{l\in J_p \cap C(i)} \varphi_l,\ \varphi_i\right) \quad \forall i \in J_p$$

and

$$\sup_{\theta\in\Theta_i} P_\theta(\varphi_i = 1) \leqslant \alpha \quad \forall i \in J_p.$$

Set

$$\overline{\varphi}_i = \min_{j\in J_p:\Theta_j\subseteq H_i} \varphi_j \quad \forall i \in I,$$

$$\phi_i = \begin{cases} \overline{\varphi}_i, & \text{if } C(i) = \varnothing; \\ \min(\max\limits_{l\in C(i)} \phi_l,\ \overline{\varphi}_i), & \text{otherwise.} \end{cases} \qquad \forall i \in I.$$

Then $\phi \in \Phi_\alpha(\mathscr{H})$ and is strongly coherent and consonant.

The proof of Theorem 5 follows directly from Theorem 4. In Theorem 5, the boundary of each individual test φ is calculated with the consonance constraint. The coherence adjustment and partitioning principle is then applied to φ to obtain $\overline{\varphi}$. The final step of another consonance adjustment is needed because the previous coherence adjustment may cause the loss of consonance. Unfortunately, it is generally not possible to have coherence and consonance constraint satisfied simultaneously in boundary calculation as we will discuss later.

Similar to Theorem 5, we can put strongly consonance constraint to the local test and come up with a strong consonant partitioning principle in analogy to the strong partitioning principle. However, due to the complexity of requiring strong coherence and strong consonance simultaneously, the next theorem is more complicated in appearance and calculation.

Theorem 6 (Strong Consonant Partitioning Principle) *Suppose that $\mathscr{H} = \{H_i : i \in I\}$ is $\cap$-closed. Let $\Theta(J_p) = \{\Theta_i : i \in J_p\}$ be the natural partition of $\cup_{i\in I} H_i$. For each $i \in J_p$ choose test φ_i for testing H_i (or Θ_i) versus K_i such that*

$$\varphi_i = \min\left(\min_{J\in\mathcal{J}(i)} \max_{l\in J} \varphi_l,\ \varphi_i \right) \quad \forall i \in J_p$$

and

$$\sup_{\theta\in\Theta_i} P_\theta(\varphi_i = 1) \leqslant \alpha \quad \forall i \in J_p.$$

Set $\overline{\phi} = \varphi$ and then adjust for coherence and consonance such that

$$\begin{aligned} \overline{\phi}_i &= \min(\min_{J\in\mathcal{J}(i)} \max_{l\in J} \overline{\phi}_l,\ \overline{\phi}_i) \qquad \text{if } \mathcal{J}(i) \neq \varnothing, \quad \forall i \in I, \\ \overline{\phi}_i &= \min_{j\in J_p:\Theta_j\subseteq H_i} \overline{\phi}_j, \qquad \forall i \in I. \end{aligned} \tag{2.2}$$

Then $\overline{\phi} \in \Phi_\alpha(\mathcal{H})$ and is strongly coherent and strongly consonant.

The proof of Theorem 6 follows directly from Theorem 3.2 of Finner and Strassburger (2002) and the definition of strong consonance. To find $\overline{\phi}$ meeting the conditions in (2.2), we can set, at step one,

$$\begin{aligned} \overline{\phi}_i^{(1,a)} &= \min(\min_{J\in\mathcal{J}(i)} \max_{l\in J} \overline{\phi}_l^{(1,a)},\ \overline{\phi}_i^{(1,a)}) \qquad \text{if } \mathcal{J}(i) \neq \varnothing, \quad \forall i \in I, \\ \overline{\phi}_i^{(1,b)} &= \min_{j\in J_p:\Theta_j\subseteq H_i} \overline{\phi}_j^{(1,a)} \qquad \forall i \in I. \end{aligned}$$

Then at step s, set $(\overline{\phi}^{(s,a)}, \overline{\phi}^{(s,b)})$ as:

$$\begin{aligned} \overline{\phi}_i^{(s,a)} &= \min(\min_{J\in\mathcal{J}(i)} \max_{l\in J} \overline{\phi}_l^{(s-1,b)},\ \overline{\phi}_i^{(s-1,b)}) \qquad \text{if } \mathcal{J}(i) \neq \varnothing \quad \forall i \in I, \\ \overline{\phi}_i^{(s,b)} &= \min_{j\in J_p:\Theta_j\subseteq H_i} \overline{\phi}_j^{(s,a)} \qquad \forall i \in I. \end{aligned}$$

The procedure stops until $\overline{\phi}$ meets the conditions in (2.2). If the number of hypotheses in $\mathcal{H}$ is finite, we can get a strongly coherent and strongly consonant test $\overline{\phi}$ by the steps described above; otherwise, a general method to find $\overline{\phi}$ still needs further research.

Remark. Bittman et al. (2009) developed a consonant multiple testing procedure in the closure principle framework and illustrated it with an example using local sum test to test two normal means greater than 0. Despite the nature of partitioning principle more powerful than closure principle, Bittman et al. (2009)'s procedure guarantees a consistency stronger than consonance and weaker than strong consonance. The rationale of their procedure is to reject an intersection hypothesis if at least one of the implied elementary hypotheses is rejected along with all the subset hypotheses in the closure of such elementary hypotheses. The critical values of their local tests are determined such that

$$\sup_{\theta\in H_i} P_\theta(\min(\max_{j\in E(i)} \min_{h\in I: H_i\subset H_h\subseteq H_j} \varphi_h,\ \varphi_i)=1)\leqslant\alpha \quad \forall i\in I.$$

Obviously their procedure is consonant and the extra constraint on the subset hypotheses in the closure of elementary hypotheses makes it better than just consonant. However, their method can not always guarantee strong consonance. A simple example is the pairwise comparison discussed earlier, for which Bittman's procedure can not always reject at least two of H_1, H_2, H_3 when H_4 is rejected.

In summary, our proposed methods have more advantages over Bittman et al. (2009)'s procedure in three ways: (1) Our procedures are combined with partitioning principle; (2) Our procedures distinguish consonance and strong consonance with different constraints; (3) Our procedures are developed in a general form without specifying the hypotheses and tests.

Remark. An ideal partitioning principle considers the coherence and consonance constraint simultaneously such that

$$\varphi_i=\min\left(\min_{j\in J_p:\Theta_j\subseteq H_i}\varphi_j,\ \min_{J\in\mathcal{J}(i)}\max_{l\in J}\varphi_l\right) \quad \forall i\in I,$$

and

$$\sup_{\theta\in\Theta_i} P_\theta(\varphi_i = 1) \leqslant \alpha, \quad \forall i \in I.$$

However, solutions to the equations above may not exist even in the simple case with two elementary hypotheses.

2.2 Example Illustration and Simulation Study

The first example serves as a detailed illustration on how to construct the consonant partitioning test. Then the second example is devoted to a simulation study which shows the actual power improvement of the proposed procedure over the closure principle.

2.2.1 Construct consonant partitioning test

Suppose we are interested in three null hypotheses H_1, H_2 and H_3. The closure of this family of hypotheses is $\mathscr{H} = \{H_1, H_2, H_3, H_4 = H_1 \cap H_2, H_5 = H_1 \cap H_3, H_6 = H_2 \cap H_3, H_7 = H_1 \cap H_2 \cap H_3\}$ (suppose the seven hypotheses are distinct from each other). Their natural partition is defined as $\Theta_i = H_i \cap (\cup_{j:H_j\subset H_i} H_j)^c$ for $i \in \{1, 2, \cdots, 7\}$. The index sets defined in (2.1) can be specified as:

$$C(1) = E(1) = \varnothing,\ C(2) = E(2) = \varnothing,\ C(3) = E(3) = \varnothing,$$

$$C(4) = E(4) = \{1, 2\},\ C(5) = E(5) = \{1, 3\},\ C(6) = E(6) = \{2, 3\},$$

$$C(7) = \{4, 5, 6\},\ E(7) = \{1, 2, 3\}.$$

Following Theorem 4, first find local level-α test $\varphi_i^{(1)}$ for each null hypothesis in $\mathscr{H}$ such that

$$\sup_{\theta\in\Theta_i} P_\theta(\varphi_i^{(1)} = 1) \leqslant \alpha \qquad \text{for } i = \{1, 2, \cdots, 7\}.$$

Notice the supremum is obtained from the natural partition Θ_i instead of H_i. Then $\overline{\varphi}^{(1)}$ is defined as

$$\begin{aligned}
\overline{\varphi}_1^{(1)} &= \min(\varphi_1^{(1)}, \varphi_4^{(1)}, \varphi_5^{(1)}, \varphi_7^{(1)}),\\
\overline{\varphi}_2^{(1)} &= \min(\varphi_2^{(1)}, \varphi_4^{(1)}, \varphi_6^{(1)}, \varphi_7^{(1)}),\\
\overline{\varphi}_3^{(1)} &= \min(\varphi_3^{(1)}, \varphi_5^{(1)}, \varphi_6^{(1)}, \varphi_7^{(1)}),\\
\overline{\varphi}_4^{(1)} &= \min(\varphi_4^{(1)}, \varphi_7^{(1)}),\\
\overline{\varphi}_5^{(1)} &= \min(\varphi_5^{(1)}, \varphi_7^{(1)}),\\
\overline{\varphi}_6^{(1)} &= \min(\varphi_6^{(1)}, \varphi_7^{(1)}),\\
\overline{\varphi}_7^{(1)} &= \varphi_7^{(1)}.
\end{aligned}$$

Using $\overline{\varphi}^{(1)}$, we can construct $\phi^{(1)}$ as follows:

$$\begin{aligned}
\phi_1^{(1)} &= \overline{\varphi}_1^{(1)} = \min(\varphi_1^{(1)}, \varphi_4^{(1)}, \varphi_5^{(1)}, \varphi_7^{(1)}),\\
\phi_2^{(1)} &= \overline{\varphi}_2^{(1)} = \min(\varphi_2^{(1)}, \varphi_4^{(1)}, \varphi_6^{(1)}, \varphi_7^{(1)}),\\
\phi_3^{(1)} &= \overline{\varphi}_3^{(1)} = \min(\varphi_3^{(1)}, \varphi_5^{(1)}, \varphi_6^{(1)}, \varphi_7^{(1)}),\\
\phi_4^{(1)} &= \min(\overline{\varphi}_4^{(1)}, \max(\phi_1^{(1)}, \phi_2^{(1)})) = \min\left(\varphi_4^{(1)}, \varphi_7^{(1)}, \max\big(\min(\varphi_1^{(1)}, \varphi_5^{(1)}), \min(\varphi_2^{(1)}, \varphi_6^{(1)})\big)\right),\\
\phi_5^{(1)} &= \min(\overline{\varphi}_5^{(1)}, \max(\phi_1^{(1)}, \phi_3^{(1)})) = \min\left(\varphi_5^{(1)}, \varphi_7^{(1)}, \max\big(\min(\varphi_1^{(1)}, \varphi_4^{(1)}), \min(\varphi_3^{(1)}, \varphi_6^{(1)})\big)\right),\\
\phi_6^{(1)} &= \min(\overline{\varphi}_6^{(1)}, \max(\phi_2^{(1)}, \phi_3^{(1)})) = \min\left(\varphi_6^{(1)}, \varphi_7^{(1)}, \max\big(\min(\varphi_2^{(1)}, \varphi_4^{(1)}), \min(\varphi_3^{(1)}, \varphi_5^{(1)})\big)\right),\\
\phi_7^{(1)} &= \min(\overline{\varphi}_7^{(1)}, \max(\phi_4^{(1)}, \phi_5^{(1)}, \phi_6^{(1)}))\\
&= \min\left(\varphi_7^{(1)}, \max\big(\min(\varphi_1^{(1)}, \varphi_4^{(1)}, \varphi_5^{(1)}), \min(\varphi_2^{(1)}, \varphi_4^{(1)}, \varphi_6^{(1)}), \min(\varphi_3^{(1)}, \varphi_5^{(1)}, \varphi_6^{(1)})\big)\right),
\end{aligned} \tag{2.3}$$

Following Theorem 5, the local test $\varphi_i^{(2)}$ for each individual hypothesis is obtained

such that

$$\begin{aligned}
&\sup_{\theta\in\Theta_1} P_\theta(\varphi_1^{(2)}=1)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_2} P_\theta(\varphi_2^{(2)}=1)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_3} P_\theta(\varphi_3^{(2)}=1)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_4} P_\theta\big(\varphi_4^{(2)}=1 \text{ and } \max(\varphi_1^{(2)},\varphi_2^{(2)})=1\big)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_5} P_\theta\big(\varphi_5^{(2)}=1 \text{ and } \max(\varphi_1^{(2)},\varphi_3^{(2)})=1\big)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_6} P_\theta\big(\varphi_6^{(2)}=1 \text{ and } \max(\varphi_2^{(2)},\varphi_3^{(2)})=1\big)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_7} P_\theta\big(\varphi_7^{(2)}=1 \text{ and } \max(\varphi_4^{(2)},\varphi_5^{(2)},\varphi_6^{(2)})=1\big)\leqslant\alpha.
\end{aligned}\tag{2.4}$$

Again the supremum is obtained under the natural partition. It is more powerful than $\varphi^{(1)}$ in Theorem 4 due to the consonance constraint in $\varphi_4^{(2)},\varphi_5^{(2)},\varphi_6^{(2)}$ and $\varphi_7^{(2)}$. Then the consonant partitioning test $\phi^{(2)}$ can be constructed similarly as (2.3) with local test $\varphi^{(1)}$ substituted by $\varphi^{(2)}$.

Applying Theorem 6, the local test is obtained such that

$$\begin{aligned}
&\sup_{\theta\in\Theta_1} P_\theta(\varphi_1^{(3)}=1)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_2} P_\theta(\varphi_2^{(3)}=1)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_3} P_\theta(\varphi_3^{(3)}=1)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_4} P_\theta\big(\varphi_4^{(3)}=1 \text{ and } \max(\varphi_1^{(3)},\varphi_2^{(3)})=1\big)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_5} P_\theta\big(\varphi_5^{(3)}=1 \text{ and } \max(\varphi_1^{(3)},\varphi_3^{(3)})=1\big)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_6} P_\theta\big(\varphi_6^{(3)}=1 \text{ and } \max(\varphi_2^{(3)},\varphi_3^{(3)})=1\big)\leqslant\alpha,\\
&\sup_{\theta\in\Theta_7} P_\theta\Big(\varphi_7^{(3)}=1 \text{ and } \max(\varphi_4^{(3)},\varphi_5^{(3)})=1 \text{ and } \max(\varphi_4^{(3)},\varphi_6^{(3)})=1 \text{ and } \max(\varphi_5^{(3)},\varphi_6^{(3)})=1\\
&\qquad \text{and } \max(\varphi_1^{(3)},\varphi_6^{(3)})=1 \text{ and } \max(\varphi_2^{(3)},\varphi_5^{(3)})=1 \text{ and } \max(\varphi_3^{(3)},\varphi_4^{(3)})=1\\
&\qquad \text{and } \max(\varphi_1^{(3)},\varphi_2^{(3)},\varphi_3^{(3)})=1\Big)\leqslant\alpha.
\end{aligned}$$

The strongly consonant partitioning test is at least as powerful as the consonant partitioning

test in (2.4) since there are more constraint on the local tests. The multiple test $\overline{\phi}^{(3)}$ can be constructed following the conditions in (2.2) of Theorem 6 with local test $\varphi^{(3)}$.

2.2.2 Simulation Study

We consider the example of pairwise comparison in Sonnemann (2008). Let $\mathbf{X}_i = (X_{i1}, \cdots, X_{in})$ denote independent normal random variables with mean θ_i and common known variance $\sigma^2 = 1$, for $i = 1, 2, 3$. Denote $\overline{X}_i$ as the sample means for $i = 1, 2, 3$ and $\overline{X}$ as the pooled sample mean. The $\cap$-closed family of null hypotheses is

$$H_1 : \theta_1 = \theta_2, \quad H_2 : \theta_1 = \theta_3, \quad H_3 : \theta_2 = \theta_3, \quad H_4 : \theta_1 = \theta_2 = \theta_3$$

In the regular closed test, the local level α test rejects H_1 if $\sqrt{n/2}|\overline{X}_1 - \overline{X}_2| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. Similar tests exist for H_2 and H_3. The intersection hypothesis H_4 is rejected if $\sum_{i=1}^{3}\sum_{j=1}^{n}(X_{ij} - \overline{X})^2 > \chi^2_{3n-1,\alpha}$, where $\chi^2_{3n-1,\alpha}$ is the $1 - \alpha$ quantile of Chi-square distribution with degrees of freedom $3n - 1$. Finally, the multiple test rejects H_i if the local test rejects H_i and H_4 for $i = 1, 2, 3$.

Following Theorem 5, we use the same local level α test for H_1, H_2, H_3, but the local test for H_4 rejects if $\sum_{i=1}^{3}\sum_{j=1}^{n}(X_{ij} - \overline{X})^2$ greater than some critical value and at least *one* of H_1, H_2, H_3 are rejected by local tests. The critical value is calculated such that the rejection of H_4 has a probability less than α when the true means are all equal. It is smaller than the critical value $\chi^2_{3n-1,\alpha}$ used by the closed test, thus the consonant partitioning test has a better chance to detect the significance of H_4. The multiple level α test then rejects H_i ($i = 1, 2, 3$) if H_i and H_4 are rejected by the local tests.

Following Theorem 6, we use the same local level α test for H_1, H_2, H_3, but the local test for H_4 rejects if $\sum_{i=1}^{3}\sum_{j=1}^{n}(X_{ij} - \overline{X})^2$ greater than some critical value and at least *two* of H_1, H_2, H_3 are rejected by local tests. The critical value is calculated such that the rejection

of H_4 has a probability less than α when the true means are all equal. It is even smaller than the critical value used in Theorem 5, thus makes the strong consonant partitioning test most powerful amongst the procedures discussed here. The multiple level α test then rejects H_i ($i = 1, 2, 3$) if H_i and H_4 are rejected by the local tests.

In the first simulation study, we artificially design $\theta_1 = -0.4$, $\theta_2 = 0$, $\theta_3 = 0.6$ such that θ_1, θ_3 have the largest difference and θ_1, θ_2 have the minimum difference. We expect that H_2 and H_4 are easier to be rejected than H_1 and H_3. An inconsistent conclusion occurs when H_4 is rejected alone. If H_4 is rejected with one of H_1, H_2, H_3, it is consonant but not strongly consonant and thus is avoided by the strong consonant partitioning test for the sake of complete consistency. With such $\boldsymbol{\theta}$ configuration and the lack of protection from consonance adjustment, the closed test is very easy to fall into the trap of inconsistency and reject H_4 only. A strong consonant conclusion includes the case that H_4 is rejected with another two or three hypotheses from H_1, H_2, H_3. The ideal case is that all four hypotheses are rejected. The (strong) consonance partitioning test is expected to have better performance and eliminates inconsistent conclusions. Finally, a failure happens when none of the hypotheses is rejected. We repeat the simulation for $100\,000$ times and the counts of each cases for the three aforementioned testing procedures are compared in Table 2.1.

As we can see from Table 2.1, inconsistent conclusions are completely eliminated by the (strong) consonant partitioning test due to the consonance adjustment step in the test procedure (Theorem 5 and 6). Moreover, by excluding the inconsistent rejections, the (strong) consonant partitioning test can reject more false null hypotheses, thus results in a lower failure rate. With sample size $n = 30$, the strong consonant partitioning test has a failure rate of only 16.9%, while the consonant partitioning test obtains a failure rate of 30.6%. Both procedures outperform the closed test with failure rate as high as 70.5%. In terms of consistency, the strong

	Inconsistent	Consonant Partial Success	Strong Consonant More Success	Strong Consonant Full Success	Failure
$(\varphi_1, \varphi_2, \varphi_3, \varphi_4) =$	$(0,0,0,1)$	$(1,0,0,1)$ $(0,1,0,1)$ $(0,0,1,1)$	$(1,1,0,1)$ $(1,0,1,1)$ $(0,1,1,1)$	$(1,1,1,1)$	$(0,0,0,0)$
Closed Test	121 (1662)	1698 (4113)	20102 (10138)	7618 (645)	70461 (83442)
Consonant Partitioning Test	0 (0)	7020 (14063)	49396 (22709)	12988 (813)	30596 (62415)
Strong Consonant Partitioning Test	0 (0)	0 (0)	68588 (33591)	14541 (833)	16871 (65576)

Table 2.1: The counts of several consistent or inconsistent rejection events for three testing procedures in the first simulation study. No other cases are possible since the procedures compared here are all coherent tests. The computation is based on $100\,000$ simulations with $\theta_1 = -0.4$, $\theta_2 = 0$, $\theta_3 = 0.6$ and $\sigma^2 = 1$. All procedures control the local test at level $\alpha = 0.05$. Sample size $n = 30$ and $n = 10$ (results showed in parentheses) are considered for each population.

consonant partitioning test results in 83.1% strongly consonant conclusions (more success and full success), while the consonant partitioning test has only 62.4% inferences strongly consonant and 7% consonant but not strongly consonant. As a result, the strong consonance shows advantage in improving consistency as well as power. When sample size is small ($n = 10$), the failure rates of all three procedures raise to some undesired level. However, both new procedures can still avoid the inconsistent conclusions. The strong consonant partitioning test can still guarantee strong consonance. As a conclusion, the two proposed procedures are superior in keeping the consistency regardless of the sample size. On the other hand, to attain the required

power, it is important to determine appropriate sample size in the experimental design stage.

In a second simulation, we let $\theta_1 = \theta_2 = 0$ and $\theta_3 = 0.8$ such that H_2, H_3 and H_4 are false null hypotheses. It differs from the first simulation in the sense that there is a true null hypothesis and the test procedures might make false rejection. If a multiple test procedure rejects H_1, it makes a familywise error (FWE). An inconsistent conclusion can be formed if H_4 is rejected alone. If H_4 is rejected with one of H_2, H_3, it is consonant but not strongly consonant. The ideal (strongly consonant) case is that all three false null hypotheses are rejected and a failure happens when none of the hypotheses is rejected. We repeat the simulation for $100\,000$ times and the counts of each cases for the three aforementioned testing procedures are compared in Table 2.2.

The result in Table 2.2 shows that all three procedures successfully control the FWER at level $\alpha = 0.05$, and the closed test has the most conservative FWER amongst the three procedures. The (strong) consonant partitioning test eliminates the possibility of inconsistency. With sample size $n = 30$, the (strong) consonant partitioning test has a clear advantage in the rate of success. With small sample size $n = 10$, the failure rates for all three procedures raise. Again we demonstrate in this simulation, the consonance and strong consonance property of the two new procedures remain intact regardless of sample size.

Certainly, the $\boldsymbol{\theta}$ configurations in the two simulation studies are carefully designed to show how awkward the closed test procedure might be. The closed test is much more conservative relative to the two new procedures due to the fact that it does not consider the consonance adjustment to sharpen the critical values, thus has the largest failure rate (fail to reject any null hypotheses). We have already known that a failure occurs if the procedure fails to reject $H_4 : \theta_1 = \theta_2 = \theta_3$. From the discussion above we know that (strong) consonant partitioning test treats H_4 differently and have a better chance to detect significance than the closed test.

This advantage comes from the consonance (strong consonance) adjustment.

2.2.3 Discussion

In this chapter, the general consonant partitioning principle (Theorem 4) is proposed in analogy to Finner and Strassburger (2002)'s weak partitioning principle. The consonant partitioning principle and strong consonant partitioning principle (Theorem 5, 6) are in analogy to the strong partitioning principle. As they have showed that the WPP and SPP can lead to more powerful procedures than the closure principle, the proposed procedures have similar advantage. However, as SPP may sometimes provide no improvement over WPP, the consonant partitioning principle and strong consonant partitioning principle can not always lead to a more powerful test than the general consonant partitioning principle or the WPP. In fact, consonance adjustment makes sense only for the class of non-consonant test procedures. Therefore the paper is motivated in the case when non-union-intersection local test is of interest, for example likelihood ratio test (Marcus et al., 1976; Tang and Lin, 1997). Comparing to Finner and Strassburger (2002)'s weak partitioning principle and strong partitioning principle, the new procedures not only ensure strong coherence, but also consonance (strong consonance). Any dissonant (but coherent) multiple level α test can be improved by a consonant one without any cost in terms of inferences on elementary hypotheses.

The three theorems outlined in this chapter provides a fairly general framework for constructing coherent and consonant multiple tests. If a multiple test procedure is already consonant, for example Dunnett's test (Dunnett, 1955), the consonant partitioning principle is equivalent to the WPP. Since any closure principle can be replaced by the partitioning principle without extra cost, the consonant partitioning principle proposed in this work is at least as good as the consonance adjustment using closure principle.

Strongly consonant procedures show advantages in terms of power and consistency

compared with consonant procedures. However, such procedures require more complicated adjustment steps and thus are more computationally expensive (Theorem 6). A proper method should be chosen according to the requirement of the research. A more specific discussion on consistent MTP in the context of multiple comparison with a control problem is discussed in the next chapter where normality is assumed for inter-treatment comparisons and the likelihood ratio test is applied as local test.

	FWE	Inconsistent	Consonant Partial Success	Strong Consonant Full Success	Failure
$(\varphi_1, \varphi_2, \varphi_3, \varphi_4) =$	$(1,0,0,1)$ $(1,1,0,1)$ $(1,0,1,1)$ $(1,1,1,1)$	$(0,0,0,1)$	$(0,1,0,1)$ $(0,0,1,1)$	$(0,1,1,1)$	$(0,0,0,0)$
Closed Test	1649 (1399)	257 (1980)	1364 (3361)	20800 (7373)	75930 (85887)
Consonant Partitioning Test	3701 (3109)	0 (0)	6216 (12376)	53126 (17394)	36957 (67121)
Strong Consonant Partitioning Test	4897 (4314)	0 (0)	0 (0)	76615 (26389)	18488 (69297)

Table 2.2: The counts of several consistent or inconsistent rejection events for three testing procedures in the second simulation study. No other cases are possible since the procedures compared here are all coherent tests. The computation is based on $100\,000$ simulations with $\theta_1 = \theta_2 = 0$, $\theta_3 = 0.8$ and $\sigma^2 = 1$. All procedures control the local test at level $\alpha = 0.05$. Sample size $n = 30$ and $n = 10$ (results showed in parentheses) are considered for each population.

Chapter 3

Consonant Closed Likelihood Ratio Test Procedures with Application to Dose-Response Study

3.1 Introduction

In the previous chapter, the consonant test is introduced under the partition testing framework. The theorems proposed in that chapter do not have any distributional assumptions and specific forms for local tests. It is a fairly general approach to construct consonant (strongly consonant) multiple testing procedures. In this chapter, a specific type of problem is considered and the multiple testing procedures proposed in this chapter takes specific forms. In this case, the consonant test can be constructed with a closed form and we can study its properties by comparing with existing multiple testing procedures. Since closure principle can be viewed as a special case of partitioning principle which takes simple form, we propose the first consonant likelihood ratio test under the closed testing framework.

Specifically, we consider the common application in clinical trials of multiple comparisons with a control as defined in 1.2.4. Label the treatments by $0, 1, 2, \ldots, k$ where 0 is the control treatment. The hypotheses $H_i : \theta_i = 0$, $i \in K = \{1, 2, \ldots, k\}$ form a family of hypotheses $\mathcal{H}$. Use the same definitions for $\bar{Y}_i$, S^2 and T_i as in (1.4) and (1.5). A closed test procedure first constructs a closure family by including all intersection hypotheses $H_I = \cap_{i \in I} H_i$ for $I \subseteq K$. Then it rejects null hypothesis H_i if all the intersection hypotheses H_I $(I \ni i)$ are rejected at individual level α. Closed test procedures satisfy the logical requirement of *coherence* introduced in Chapter 2, which says, if H_1 is retained and $H_1 \subset H_2$, then H_2 can not be rejected.

Different choice of individual level α tests for the intersection hypotheses H_I results in various closed test procedures. Using the normality assumption as described above, the step-down Dunnett test (SD) and Dunnett-Tamhane step-up test (SU) are powerful choices, which have been introduced in Section 1.2.4 and 1.2.5. Without using the normality assumptions, Bonferroni procedures such as Holm test, Hochberg test and Hommel test can be good alternatives. Notice that both Hochberg test and Hommel test rely on the assumption of independent test statistics and are at least as powerful as Holm's step-down test, which does not need such assumptions. For more detailed introduction and comparisons of aforementioned closed test procedures, one can refer to Tamhane (1996); Huang and Hsu (2007); Dunnett and Tamhane (1993) and Hsu (1996).

It has been shown in Chapter 1 that the closed testing procedures in the literature are all based on the union-intersection method. An alternative to the UI method is to test the intersection hypotheses by likelihood ratio test. Likelihood ratio test is not widely used in closed testing procedures since it does not allow step-wise shortcut, thus the procedure is slightly more complex and may require more computation time. However, the modern computer technology has dramatically shortened the difference in computation time. Both step-wise and non-step-

wise testing procedures can be done in mere seconds. As a result, the power advantage of likelihood ratio test in certain cases is of more interest.

To construct a powerful closed test procedure based on likelihood ratio test, we need to emphasize the logical constraint of consonance, which requires that if any hypothesis is rejected then at least one hypothesis implied by it must also be rejected. Specifically, if an intersection hypothesis $H_I = \cap_{i \in I} H_i$ is rejected, then at least one of the component hypotheses H_i must be rejected.

A common agreement is that closed test procedures are coherent by definition but not necessarily consonant. We would like to find out how to change a dissonant closed test into a consonant one and whether such adjustment is helpful. In this chapter, we first show that two powerful closed test procedures (SD test and SU test) are consonant. Then we will show that likelihood ratio based closed tests are not consonant and how to add consonance adjustment to them. As we have shown in Section 2.2.2, adding consonance adjustment to dissonant tests can improve the test power by excluding inconsistent rejections. Hence a consonant closed test is more powerful than its dissonant counterpart. Moreover, being both coherent and consonant, the consonant closed likelihood ratio test becomes a powerful competitor to the SD and SU test. The advantage comes from the optimality of likelihood ratio test.

Bittman et al. (2009) constructed a consonant closed test procedure using sum test for the comparisons with a control problem assuming known variance. They also showed that such consonant closed tests are optimal maximin test. We focus on the comparisons with a control problem with unknown variance and propose consonant closed test procedures based on likelihood ratio test. This procedure can be viewed as a special case of Theorem 4 with closed form and shortcut. We will show by simulation and dose-response study examples that the consonant closed likelihood ratio tests have optimal performance credited to the shape of their rejection region. We also show that the consonant closed likelihood ratio tests do not rely

on the assumption of independent or equally correlated test statistics, thus is preferable to SU, Hochberg and Hommel test.

In Section 3.2, The step-down Dunnett test (SD) and Dunnett-Tamhane step-up test (SU) are illustrated as benchmarks to compare with the proposed consonant closed likelihood ratio tests. In Section 3.3, We introduce consonance adjustment in the simple case where the likelihood ratio test is equivalent to F test. The rejection region and performance of the consonant closed test is compared with SD test and SU test. In Section 3.4, We extend the discussion to one-sided alternative problem which involves order restricted test statistics. Applications to dose-response data are presented in Section 3.5. In Section 3.6.1, We extend the consonant closed likelihood ratio test to multivariate analysis where longitudinal data presented in the model. The computation issue is discussed in Section 3.6.2.

3.2 Step-down Dunnett (SD) Test and Dunnett-Tamhane Step-up (SU) Test as Benchmarks

The procedure of SD and SU tests have been reviewed in Section 1.2.4 and 1.2.5. For the SD test, it is easy to show that the critical constants $t_1, t_2, \ldots, t_k$ are monotonically increasing. This monotonicity suggests that the SD test is consonant.

Using the same notations as Section 1.2.4, for any intersection hypothesis H_I, suppose

$$H_I = \cap_{i \in I} H_i \subset H_J = \cap_{j \in J} H_j,$$

where J is any subset of I. Consonance requires that if H_I is rejected then at least one of H_J is also rejected. Suppose the sample sizes are equal among the treatments.

The SD test will reject H_I if

$$\max_{i \in I} T_i > t_{|I|},$$

in the case of equal sample size in analogous to (1.6). Obviously there exists a subset $J \subset I$ such that $\max_{j \in J} T_j = \max_{i \in I} T_i > t_{|I|} > t_{|J|}$ as implied by the monotonicity of critical constants. Therefore the SD test rejects at least some H_J for $J \subset I$ and it is consonant.

In the case of $k = 2$, $n_1 = n_2 = 30$ and $n_0 = 40$, we can calculate the critical constants as $t_1 = 1.661$ and $t_2 = 1.948$ for $\alpha = 0.05$. At the first step, the SD test rejects the hypothesis $H_{(2)}$ if $\max(T_1, T_2) > 1.948$ and go to the second step. Otherwise, if $\max(T_1, T_2) \leqslant 1.948$, that is both T_1 and T_2 not larger than 1.948, both hypotheses are retained. At the second step, the SD test rejects $H_{(1)}$ if $T_{(1)} > 1.661$, otherwise retain $H_{(1)}$.

The step-wise procedure can be viewed from another perspective. As the condition $\max(T_1, T_2) > 1.948$ must be satisfied in order to reject any hypothesis. H_1 should be rejected if $T_1 > 1.661$ and $\max(T_1, T_2) > 1.948$. H_2 should be rejected if $T_2 > 1.661$ and $\max(T_1, T_2) > 1.948$. The resulting rejection region is displayed in Figure 3.1. The area labeled '1' is the region that both null hypotheses $H_1 : \theta_1 = 0$ and $H_2 : \theta_2 = 0$ are to be rejected. The area labeled '2' is the region that only one of the two hypotheses will be rejected. The area labeled '3' is the region to retain both hypotheses.

The SU test has been reviewed in Section 1.2.5, which compares the ordered t statistics with critical constants satisfying the equations in (1.9). However, no analytical proof is available so far to show that solutions $c_1, \ldots, c_m$ satisfying the monotonicity conditions exist for arbitrary m when T_i's are correlated (through the common variance estimator and the control treatment sample mean). Suppose this conjecture is true, the SU test is consonant as well. The proof is provided here.

For any intersection hypothesis H_I, the SU test will reject H_I if (1.8) is satisfied. Specifically, there exists at least one index $d \in I$ with rank r among the t statistics in I, such that $T_d = T_{(r)} > c_r$ is satisfied. For any subset $J \subset I$ and $J \ni d$, denote $T_{(1)^J} \leqslant T_{(2)^J} \leqslant \ldots \leqslant$

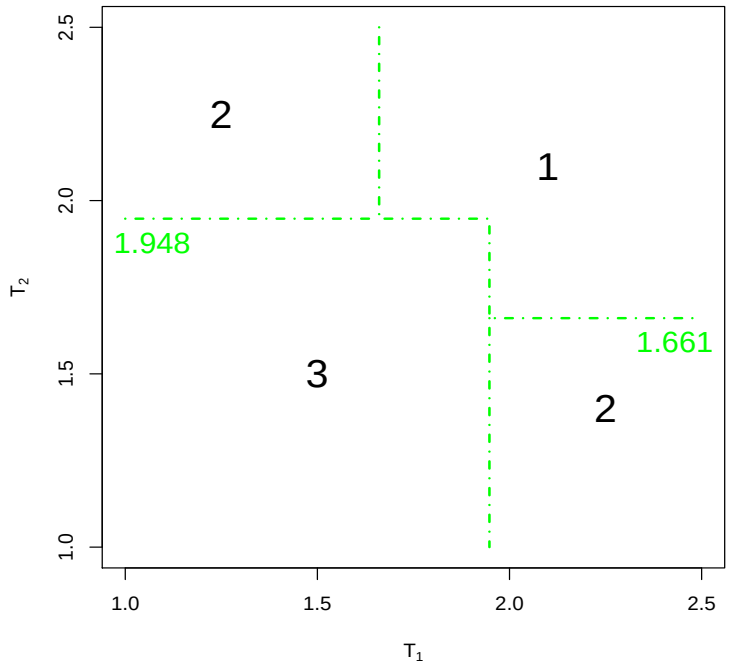

Figure 3.1: The rejection region for SD test. '1' is the region to reject both hypotheses; '2' is the region to reject one but not both hypotheses; '3' is the region to retain both hypotheses.

$T_{(|J|)^J}$ as the ordered t statistics among J. It is obvious that the new rank of T_d is at most r, that is $T_d \leqslant T_{(r)^J}$. Due to the monotonicity of critical constants, $T_d > c_r \geqslant \ldots \geqslant c_1$. Therefore the SU test reject at least some H_J for $J \subset I$ and it is consonant.

In the case of $k = 2$, $n_1 = n_2 = 30$ and $n_0 = 40$, we use the algorithm introduced by Kwong and Liu (2000) to calculate the critical constants as $c_1 = 1.661$ and $c_2 = 1.963$ for $\alpha = 0.05$. At the first step, the SU test retains the hypothesis $H_{(1)}$ if $\min(T_1, T_2) \leqslant 1.661$ and go to the second step. Otherwise, if $\min(T_1, T_2) > 1.661$, that is both T_1 and T_2 larger than 1.661, both hypotheses are rejected. At the second step, the SU test retains $H_{(2)}$ if $\max(T_1, T_2) \leqslant 1.963$, that is both T_1 and T_2 less than 1.963. Otherwise rejects $H_{(2)}$.

In another perspective, reject both hypotheses if $T_1 > 1.661$ and $T_2 > 1.661$. Reject one of the two hypotheses if $T_{(1)} \leqslant 1.661$ and $T_{(2)} > 1.963$ and retain both hypotheses if

$T_{(1)} \leqslant 1.661$ and $T_{(2)} \leqslant 1.963$. The resulting rejection region is displayed in Figure 3.2.

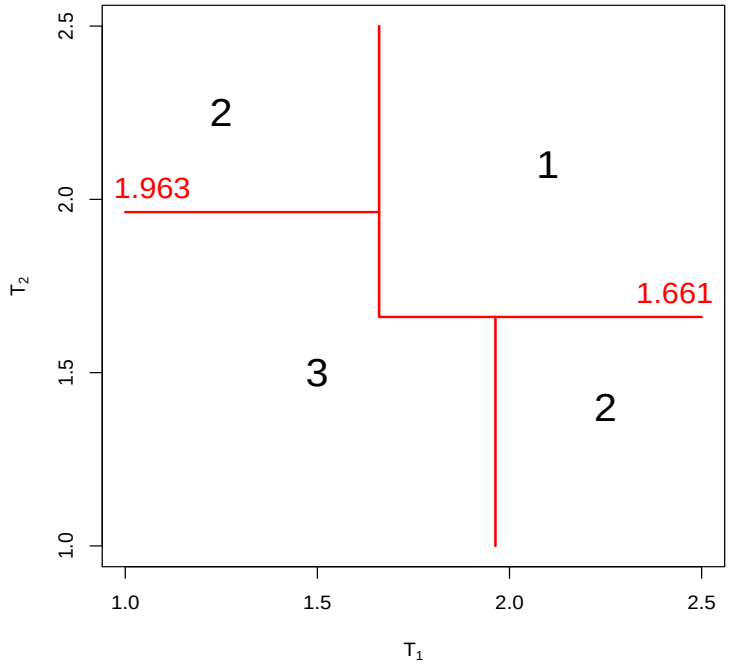

Figure 3.2: The rejection region for SU test. '1' is the region to reject both hypotheses; '2' is the region to reject one but not both hypotheses; '3' is the region to retain both hypotheses.

Compare the rejection regions in Figure 3.1 and 3.2, we see that neither of the two test procedures dominates the other. This conclusion is proved in Dunnett and Tamhane (1992) by showing $c_i > t_i$ for $i \geqslant 2$. However, if most hypotheses are false then the SU test is moderately more powerful than the SD test Dunnett and Tamhane (1993).

These two step-wise procedures are compared later with the new consonant closed likelihood ratio test as benchmarks. As they all use the normality assumption and are more powerful than the Bonferroni version of step-wise tests introduced in Chapter 1.

3.3 Consonant closed F test against two-sided alternatives

3.3.1 A consonant closed F test in the simple case

Now we explore the alternative choice of applying likelihood ratio test to intersection hypotheses in closed test procedures. To keep things simple at the beginning, we first consider a dose-response study comparing 2 doses of an experimental drug (labeled '1' and '2') to a control drug (placebo or an active control, labeled '0') with two-sided alternative hypothesis, $H_i : \theta_i = 0$ vs. $A_i : \theta_i \neq 0$ $(i = 1, 2)$. The test statistics for H_1 and H_2 are $|T_1|$ and $|T_2|$ as the absolute value of the t statistics defined in (1.5). The intersection hypothesis is $H_{12} : \theta_1 = \theta_2 = 0$ against the alternative that at least one of θ_1 and θ_2 is not 0. The likelihood ratio test statistic

$$F = \frac{\sum_{i=0}^{2} n_i(\bar{Y}_i - \bar{Y})^2/2}{S^2} \tag{3.1}$$

follows an F distribution, where $\bar{Y} = \sum_{i=0}^{2} n_i\bar{Y}_i / \sum_{i=0}^{2} n_i$ is the pooled mean estimator and S^2 is the unbiased mean squared error estimator of σ^2.

The closed test procedure rejects H_i $(i = 1, 2)$ if $|T_i| > t_{\nu,\alpha/2}$ and $F > F_{2,\nu,\alpha}$, where $t_{\nu,\alpha/2}$ is the upper $\alpha/2$ quantile of t distribution with $\nu = n_0 + n_1 + n_2 - 3$ degrees of freedom and $F_{2,\nu,\alpha}$ is the upper α quantile of F distribution with degrees of freedom 2 and ν. It is not a consonant test since the rejection of H_{12} does not imply the rejection of H_1 or H_2.

We propose to add a consonance adjustment step to the F test of the intersection hypothesis, such that H_{12} is rejected if the statistic F in (3.1) exceeds some critical constant f and at least one of H_i is rejected by $|T_i| > t_{\nu,\alpha/2}$ to guarantee consonance. Such critical constant f can be found by solving the following equation:

$$P\big(F > f \text{ and } |T_i| > t_{\nu,\alpha/2} \text{ for } i = 1 \text{ or } 2\big) = \alpha.$$

Then the resulting consonant closed F test rejects H_i if $|T_i| > t_{\nu,\alpha/2}$ and $F > f$, for $i = 1, 2$.

Since a closed form of f does not exist, we solve it by Monte-Carlo simulation as the upper α quantile of a truncated F distribution. The truncation works as follows: First, generate m sets of normal random samples following the overall null hypothesis H_{12}; Then calculate F and $|T_i|$'s for each set of generated samples; If $|T_i| \leqslant t_{\nu,\alpha/2}$ for both $i = 1$ and 2, set $F = 0$; Repeat the truncation for each of the m simulated samples and estimate the upper α quantile using one of the quantile estimation methods discussed in Hyndman and Fan (1996). It is obvious that f is less than or equal to $F_{2,\nu,\alpha}$ due to the truncation. As a result, the consonant closed F test is uniformly more powerful than the unadjusted test with the family-wise error rate controlled at the same level.

The p-value approach follows similar steps. Suppose $\{y_{ij}^0 : i = 0, 1, 2;\ j = 1, \ldots, n_i\}$ is the observed responses from all subjects in 2 treatment groups and the control group. Denote the observed test statistics by f^0, $|t_1^0|$ and $|t_2^0|$. The p-value is calculated by Monte-Carlo integration which generates m sets of normal random samples following the overall null hypothesis. Calculate F and $|T_i|$'s for each set of generated samples; If $|T_i| \leqslant t_{\nu,\alpha/2}$ for both $i = 1$ and 2, set $F = 0$. If both $|t_1^0|$ and $|t_2^0|$ are less than $t_{\nu,\alpha/2}$, set p-value$= \alpha$ and fail to reject the null hypothesis H_{12}. Otherwise, estimate p-value by the proportion of m samples that has $F > f^0$. More discussion about the accuracy of calculating the critical constants and p-values can be found in Section 3.6.2.

3.3.2 Compare with SD test and SU test

We compare the consonant closed F (CCF) test with the SD and SU test in the case of 2 treatments and a control. Suppose the sample sizes are 30 for the treatment groups and 40 for the control group. For significant level $\alpha = 0.05$, the critical constants of the CCF test is calculated as $t_{\nu,\alpha/2} = 1.985$ and $f = 2.911$ with $\nu = 97$. The critical constants for SD and SU test are calculated following the algorithms reviewed in Section 3.2 in a two-sided fashion,

which are $|t_1| = |c_1| = 1.985$, $|t_2| = 2.252$ and $|c_2| = 2.263$. For a fixed variance estimate $S^2 = 1$, the rejection regions of CCF, SU and SD tests are visualized in Figure 3.3 in terms of the t statistics (T_1, T_2). Each testing procedure cut the plane into 9 pieces. If the test statistics (T_1, T_2) fall into areas labeled '1', both null hypotheses H_1 and H_2 will be rejected, indicating both treatments are different from the control. If the test statistics fall into areas labeled '2', then one treatment is significantly different from the control but the other treatment does not show significant difference. If the test statistics fall into the area labeled '3', both treatments are not significantly different from the control. To compare, we refer to those areas as 'CCF-L', 'SD-L' and 'SU-L' with 'L' indicating the label ($L = 1, 2, 3$). In the first and third quadrant, we have SU-1 $\approx$ CCF-1 $>$ SD-1. Also notice that the difference between SU-1 and CCF-1 is much smaller than their differences to SD-1. Thus both CCF and SU tests are more powerful than SD test when most of the null hypotheses are false. In the second and fourth quadrant, CCF-1 $=$ SU-1 $>$ SD-1 and CCF-2 $\gg$ SD-2 $>$ SU-2. Hence CCF is more powerful than SU and SD test in the second and fourth quadrant. CCF-3 is smaller than both SD-3 and SU-3. To conclude, none of the procedure is uniformly more powerful than the others in all four quadrants. However, CCF test is more powerful than SD test most of the time except on four very narrow areas in the first and third quadrant. CCF test also dominates SU test more frequently than being dominated by SU test.

We also compare the procedures in the following simulation study. The observations are generated with true mean $\theta_1 \neq 0$, $\theta_2 \neq 0$ and common variance $\sigma^2 = 1$. We expect the test to reject both null hypotheses. The power in this simulation study is defined as the number of rejections among all s simulated sets of observations divided by $2s$, which is essentially the average power of correctly rejecting a false null hypothesis. The result is displayed in Figure 3.4. Several different (θ_1, θ_2) combinations are evaluated with $s = 10^5$ simulations for each

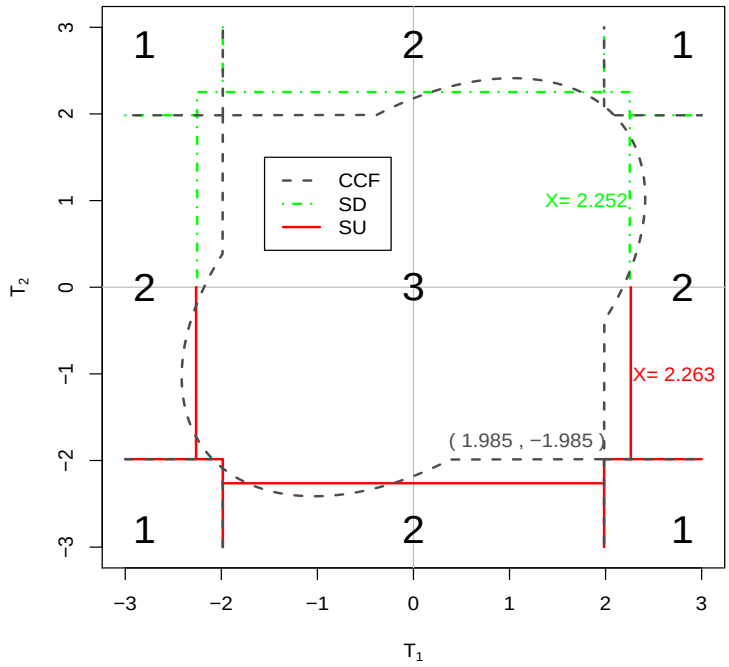

Figure 3.3: The rejection region for CCF test, SD test (upper half plane only) and SU test (lower half plane only). '1' is the region to reject both hypotheses; '2' is the region to reject one but not both hypotheses; '3' is the region to retain both hypotheses.

combination. Both θ_1 and θ_2 varies from -1 to 1 which will result in a t statistic covered roughly in the rejection region in Figure 3.3 ($1/\sqrt{1/30 + 1/40} = 4.14$). As the rejection region is symmetric with respect to the origin, we displayed the power gain of CCF test over SD test on the upper half plane (1st and 2nd quadrant), and the power gain of CCF test over SU test on the lower half plane (3rd and 4th quadrant). Since the sample variance S^2 is random instead of being fixed at 1, the contour plot does not follow exactly as the rejection region. The peak values of the contour plot appear in the 2nd and 4th quadrant where CCF test has largest rejection region (CCF-1 and CCF-2 are the largest amongst the others in the left panel). The 'hottest' areas are around $(-0.4, 0.4)$ and $(0.4, -0.4)$, which correspond to the values $(-1.66, 1.66)$ and $(1.66, -1.66)$ in the left panel (multiplied by the common factor $1/\sqrt{1/30 + 1/40}$ according to

equation (1.5)). Overall, in most areas of the contour plot (including areas beyond the display), all three tests perform similarly. In some small areas (mostly in the first and third quadrant), SU/SD tests have a power advantage over CCF test of less than 1%. On the other hand, in some larger areas in the second and fourth quadrant, CCF test has a power advantage over SU and SD tests for $4 \sim 7\%$. The advantage of CCF test is much stronger than its disadvantage, thus it is considered to be the optimal procedure.

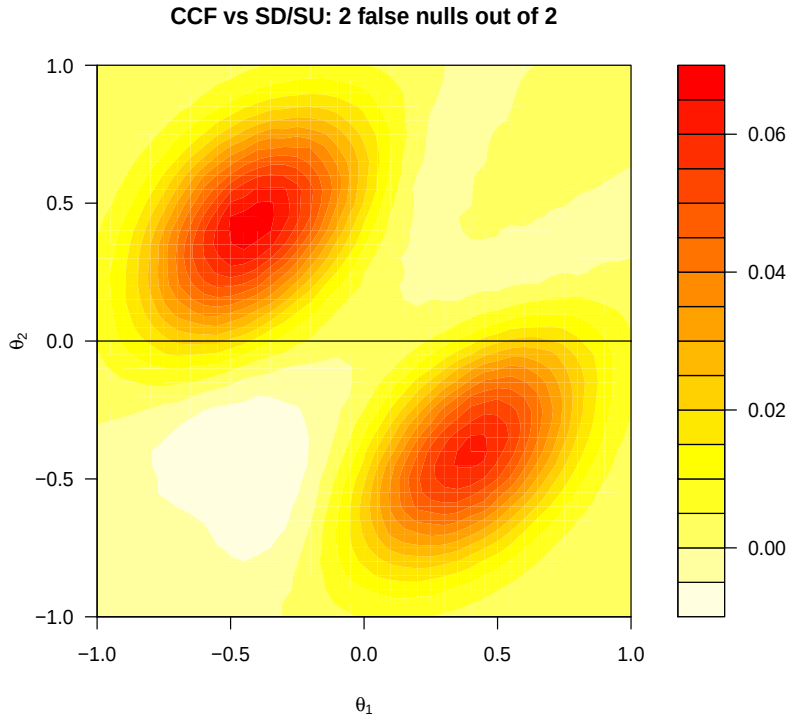

Figure 3.4: The power difference between CCF test and SD test (upper half plane) and SU test (lower half plane) for different (θ_1, θ_2) combinations. A positive value indicates the power advantage of CCF test. The result is based on 10^5 simulations.

3.3.3 A consonant closed F test comparing k doses with a control

Now we extend the CCF test to k doses. Again, suppose we are interested in testing the null hypotheses of no treatment effect against two-sided alternatives, $H_i : \theta_i = 0$ vs $A_i : \theta_i \neq 0$ for $i = 1, \ldots, k$. For any index set $I \subseteq K = \{1, \ldots, k\}$ and $|I| > 1$, the intersection hypothesis $H_I = \cap_{i \in I} H_i$ can be tested using F statistic

$$F_I = \frac{\sum_{i \in I \cup \{0\}} n_i (\bar{Y}_i - \bar{Y}_I)^2 / |I|}{S^2},$$

where $\bar{Y}_i$ is the sample mean and $\bar{Y}_I$ is the pooled sample mean of treatment $i \in I$ and the control. As usual, S^2 is the unbiased mean squared error estimator of the common variance with ν degrees of freedom.

Denote $\{\psi_I : I \subseteq K, |I| > 1\}$ as the set of local level α tests for all intersection hypotheses. The consonance adjustment is made by rejecting H_I ($\psi_I = 1$) if $F_I > f_I$ and at least one of H_J is rejected ($\psi_J = 1$), where $J \subset I$ and $|J| = |I| - 1$. The critical constant f_I satisfies

$$P\big(F_I > f_I, \max_{J \subset I, |J| = |I| - 1} \psi_J = 1\big) = \alpha, \tag{3.2}$$

which can be solved by Monte-Carlo integration recursively.

This consonance adjustment guarantees that if any k-term intersection hypothesis is rejected, then at least one of the $(k-1)$-term intersection hypotheses implied by it is also rejected. Since the adjustment is applied stepwise, it also implies that at least one rejection occurs among $(k-2)$-term intersection hypotheses, $(k-3)$-term intersection hypotheses, etc. This adjustment also guarantees that the procedure is consonant at each step. Bittman et al. (2009) proposed a procedure assuming known variance and demonstrated it in the case of comparing $k = 2$ treatments with a control. The generalization of their procedure to the unknown variance case and $k > 2$ treatments can be potentially more powerful than the CCF test but no computing algorithm is provided so far.

Incorporate all consonance adjustment steps, the resulting CCF test rejects H_i if all the H_I's with $I \ni i$ are rejected by $F_I > f_I$ and at least one rejection among H_J, where $J \subset I$ and $|J| = |I| - 1$. A shortcut exists which rejects the hypothesis $H_i : \theta_i = 0$ if $|T_i| > t_{\nu,\alpha/2}$ and for all the index set $I \subseteq K$, $|I| > 1$ and $I \ni i$, $F_I > f_I$. The validity of this shortcut is proved as follows.

Suppose $\{\psi_I : I \subseteq K, |I| > 1\}$ is the set of local level α tests for all intersection hypotheses. We denote the event of $\{\psi_I = 1\}$ by R_I and the event of $\{F_I > f_I\}$ by B_I. Then it is obvious that

$$R_I = B_I \cap \big(\cup_{J\subset I, |J|+1=|I|} R_J\big).$$

If we further notice that if I contains only one element, namely $I = \{i\}$, then B_I and R_I denote the same event of $\{|T_i| > t_{\nu,\alpha/2}\}$. Thus for any two subsets J and K, $i \in J \subset K$ and $|J| + 1 = |K|$,

$$R_K \cap R_J = \Big[B_K \cap \big(\bigcup_{\substack{L\subset K \\ |L|+1=|K|}} R_L\big)\Big] \cap R_J = B_K \cap \Big[\bigcup_{\substack{L\subset K \\ |L|+1=|K|}} \big(R_L \cap R_J\big)\Big] = B_K \cap R_J.$$

Applying this equality repeatedly, we have $\cap_{I\ni i} R_I = \cap_{I\ni i} B_I$. Thus the shortcut CCF test rejects H_i if all the H_I's with $I \ni i$ are rejected with $F_I > f_I$.

We see that there is no assumptions for equal sample sizes among treatment groups. The CCF test controls the FWER strongly in the case with unequally correlated test statistics. This property makes the CCF test preferable to the SU test which is proved to control FWER strongly only in the equal correlation case. When the sample sizes of all treatment groups are equal ($n_i = n$ for all $i = 1, \ldots, k$), the correlations among t statistics are all equal to

$$\rho = \frac{n}{n + n_0}.$$

For selected ρ and degrees of freedom, the critical constants of CCF test is reported in Table 3.1

ρ	k=2			k=3				k=4				
	df	$f_1(=t_1)$	f_2	df	f_1	f_2	f_3	df	f_1	f_2	f_3	f_4
0.1	19	2.093	3.452	20	2.086	3.430	3.075	21	2.080	3.399	3.047	2.823
	52	2.007	3.105	56	2.003	3.087	2.737	60	2.000	3.085	2.727	2.508
	85	1.988	3.032	92	1.986	3.022	2.667	99	1.984	3.013	2.654	2.439
	107	1.982	3.009	116	1.981	2.995	2.646	125	1.979	3.006	2.649	2.427
	129	1.979	2.991	140	1.977	2.992	2.634	151	1.976	2.977	2.623	2.406
	∞	1.960	2.786	∞	1.960	2.786	2.408	∞	1.960	2.786	2.408	2.210
0.3	23	2.069	3.293	28	2.048	3.204	2.839	33	2.035	3.158	2.796	2.590
	62	1.999	3.032	76	1.992	2.991	2.625	90	1.987	2.972	2.608	2.404
	101	1.984	2.979	124	1.979	2.951	2.582	147	1.976	2.932	2.569	2.367
	127	1.979	2.957	156	1.975	2.939	2.572	185	1.973	2.922	2.557	2.344
	153	1.976	2.939	188	1.973	2.925	2.560	223	1.971	2.916	2.548	2.343
	∞	1.960	2.789	∞	1.960	2.789	2.413	∞	1.960	2.789	2.413	2.211
0.5	27	2.052	3.120	36	2.028	3.024	2.654	45	2.014	2.969	2.599	2.407
	72	1.993	2.898	96	1.985	2.878	2.496	120	1.980	2.852	2.476	2.281
	117	1.980	2.861	156	1.975	2.850	2.465	195	1.972	2.840	2.455	2.254
	147	1.976	2.843	196	1.972	2.823	2.448	245	1.970	2.824	2.444	2.245
	177	1.973	2.827	236	1.970	2.818	2.440	295	1.968	2.820	2.440	2.238
	∞	1.960	2.783	∞	1.960	2.783	2.398	∞	1.960	2.783	2.398	2.204

Table 3.1: Critical constants of CCF test for selected correlation ρ and degrees of freedom, $\alpha = 0.05$.

for up to $k = 4$ treatment groups. The degrees of freedom are selected such that the corresponding sample sizes are integers.

3.4 Consonant closed order restricted test against one-sided alternatives

3.4.1 Consonant closed order restricted test

In Section 3.3, we deal with dose-response hypotheses of no treatment effect against two-sided alternatives, namely,

$$H_i : \theta_i = 0 \quad \text{vs.} \quad A_i : \theta_i \neq 0, \quad i = 1, 2, \ldots, k.$$

Their intersection hypotheses and alternatives are of the form

$$H_I : \theta_{i_1} = \theta_{i_2} = \cdots = 0 \quad \text{vs.} \quad A_I : \text{at least one of } \theta_{i_j} \neq 0 \text{ for } i_j \in I.$$

However, people are sometimes more interested in testing the null hypotheses against the one-sided alternatives in dose-response studies. In particular, hypotheses of the type

$$H_i : \theta_i \leqslant 0 \quad \text{vs.} \quad A_i : \theta_i > 0,$$

with $H_i \cup A_i$ cover all possible real values of θ_i, $i = 1, \ldots, k$. The intersection hypotheses and alternatives are of the form

$$H_I : \theta_j \leqslant 0 \text{ for all } j \in I \quad \text{vs.} \quad A_I : \theta_j > 0 \text{ for at lease one of } j \in I.$$

In this case, the likelihood ratio test statistic for H_I is

$$S_I = \frac{\sum_{i \in I \cup \{0\}} n_i (\bar{Y}_i - \tilde{Y}_i^I)^2 / |I|}{S^2},$$

where for any $i \in I \cup \{0\}$, $\bar{Y}_i$ is the sample mean of ith treatment response ($i = 0$ for control group) and $\tilde{Y}_i^I$ is the constrained maximum likelihood estimator of μ_i under the null hypothesis

$H_I : \theta_j = \mu_j - \mu_0 \leqslant 0$ for all $j \in I$. This type of likelihood ratio test is order restricted. A comprehensive introduction on order restricted inferences including algorithms to calculate the constrained MLE of μ_i under H_I can be found in Robertson et al. (1988).

Denote $\{\psi_I : I \subseteq K, |I| > 1\}$ as the set of local level α tests for all intersection hypotheses. The consonant closed order restricted test (CCOR test) rejects H_I ($\psi_I = 1$) if $S_I > s_I$ and at least one of H_J is rejected ($\psi_J = 1$), where $J \subset I$ and $|J| = |I| - 1$. The critical constant s_I satisfies the following equation

$$P\big(S_I > s_I, \max_{J \subset I, |J| = |I|-1} \psi_J = 1\big) = \alpha. \tag{3.3}$$

Since $\tilde{Y}_i^I$ is a function of $\{\bar{Y}_i : i \in I \text{ or } i = 0\}$ Robertson et al. (1988), the critical constants and p-values can be calculated recursively following the same Monte-Carlo method discussed in Section 3.3.1. Apply a similar shortcut, the CCOR test rejects H_i if $t_i > t_{\nu,\alpha}$ and for all the index sets $I \subseteq K$, $|I| > 1$ and $I \ni i$, $S_I > s_I$.

Again the CCOR test does not depend on the assumption of equally correlated test statistics, thus is superior to SU test. In the case of equal correlations, the critical constant s_I of CCOR test is reported in Table 3.2 for selected degrees of freedom.

3.4.2 Compare with SD test and SU test

First we compare the CCOR test with SD and SU test in the case of 2 treatments and a control. Suppose the sample sizes are 30 for the treatment groups and 40 for the control group. For significant level $\alpha = 0.05$, the critical constants of the CCOR test are $s_1 = t_{\nu,\alpha} = 1.661$ and $s_2 = 1.988$ with $\nu = 97$. The critical constants for SD and SU tests are $t_1 = c_1 = 1.661$, $t_2 = 1.948$ and $c_2 = 1.963$. For a fixed variance estimate $S^2 = 1$, the likelihood ratio statistic $S_{\{1,2\}}$ is determined by t statistics (T_1, T_2) and the rejection regions for CCOR, SU and SD tests are visualized in Figure 3.5. If (T_1, T_2) fall into areas labeled '1', both null hypotheses

ρ	k=2			k=3				k=4				
	df	$s_1(=t_1)$	s_2	df	s_1	s_2	s_3	df	s_1	s_2	s_3	s_4
0.1	19	1.729	2.341	20	1.725	2.329	1.993	21	1.721	2.320	1.989	1.773
	52	1.675	2.161	56	1.673	2.157	1.833	60	1.671	2.145	1.829	1.620
	85	1.663	2.125	92	1.662	2.125	1.802	99	1.660	2.112	1.793	1.590
	107	1.659	2.115	116	1.658	2.102	1.789	125	1.657	2.103	1.785	1.579
	129	1.657	2.109	140	1.656	2.103	1.781	151	1.655	2.097	1.777	1.572
	∞	1.645	1.904	∞	1.645	1.904	1.512	∞	1.645	1.904	1.512	1.268
0.3	23	1.714	2.225	28	1.701	2.174	1.791	33	1.692	2.146	1.766	1.525
	62	1.670	2.069	76	1.665	2.059	1.688	90	1.662	2.051	1.680	1.447
	101	1.660	2.038	124	1.657	2.041	1.671	147	1.655	2.027	1.661	1.426
	127	1.657	2.033	156	1.655	2.028	1.665	185	1.653	2.027	1.658	1.427
	153	1.655	2.034	188	1.653	2.019	1.658	223	1.652	2.020	1.654	1.421
	∞	1.645	1.908	∞	1.645	1.908	1.510	∞	1.645	1.908	1.510	1.269
0.5	27	1.703	2.074	36	1.688	2.026	1.613	45	1.679	2.004	1.596	1.343
	72	1.666	1.966	96	1.661	1.952	1.551	120	1.658	1.945	1.543	1.298
	117	1.658	1.947	156	1.655	1.936	1.538	195	1.653	1.927	1.533	1.287
	147	1.655	1.938	196	1.653	1.932	1.533	245	1.651	1.929	1.534	1.285
	177	1.654	1.928	236	1.651	1.928	1.530	295	1.650	1.926	1.529	1.283
	∞	1.645	1.907	∞	1.645	1.907	1.516	∞	1.645	1.907	1.516	1.275

Table 3.2: Critical constants of CCOR test for selected correlation ρ and degrees of freedom, $\alpha = 0.05$.

H_1 and H_2 will be rejected, indicating both treatments are superior to the control. If the test statistics fall into areas labeled '2', then one treatment is significantly superior to the control but the other treatment is not. If the test statistics fall into the area labeled '3', both treatments are not significantly superior to the control. Following similar notations as Section 3.3.2, SU-1 $\approx$ CCOR-1 $>$ SD-1 and they all share part of the straight boundary 1.661. CCOR-3 has a smooth shape and are relatively smaller in size. Though SD-3 and SU-3 have slight advantage on the edge of 1.948 and 1.963, they have larger disadvantage in the corner of the rectangle shaped region. The same result is obtained if we focus on regions labeled '2'. The following two simulation studies further demonstrate that the CCOR test works as good as the SD/SU test most of the time and outperform them in certain cases.

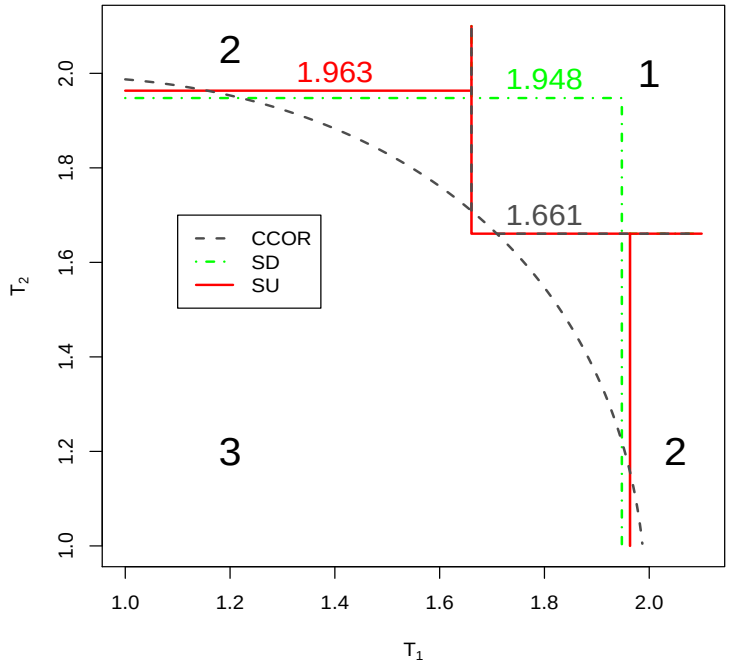

Figure 3.5: The rejection region for CCOR, SD and SU test. '1' is the region to reject both hypotheses; '2' is the region to reject one but not both hypotheses; '3' is the region to retain both hypotheses.

In the first simulation study, the observations are generated with true means $\theta_1 > 0$, $\theta_2 > 0$ and common variance $\sigma^2 = 1$. We expect to reject both null hypotheses. The power in this simulation study is defined as the number of rejections among all s simulated sets of observations divided by $2s$. The result is displayed in Figure 3.6 and 3.7 for $s = 10^5$. Figure 3.6 shows the power difference between CCOR test and SD test and a positive value indicates the power advantage of CCOR test. 3.7 shows the power difference between CCOR test and SU test. Both graphs exploit (θ_1, θ_2) combinations vary from 0 to 1 which are roughly the same region covered in Figure 3.5. The contour plots differ from the rejection regions in Figure 3.5 since the sample variance S^2 varies for each simulation (instead of being fixed at 1). The power advantage of CCOR test over SD test peaks roughly at $\theta_1 = \theta_2 = 0.45$ (t statistics around $0.45/\sqrt{1/30 + 1/40} = 1.86$ in Figure 3.5) which is inside the rejection region of CCOR test but not SD test. In most areas of 3.6, CCOR test and SD test perform similarly. In some small area SD test has a power advantage of about 0.5%, while the power of CCOR test can exceed that of SD test by 1.7% in a larger area. To compare with SU test, we see that most of the time their difference is within $\pm 0.2\%$. In the middle of the right panel, the power of CCOR test is 0.6% higher than SU test. Although the power gain of CCOR test is marginal in this simulation since SD and SU tests are already very powerful, we still prefer CCOR test since it works at least as good as SD/SU tests for any parameter configurations. Also bear in mind that CCOR does not depend on the equal correlation assumption as the SU test.

To consider different aspects of power, a more thorough evaluation is presented in Table 3.3. The performance of CCOR, SD and SU tests are evaluated at selected $\boldsymbol{\theta}$ combinations where both null hypotheses are false. We measure the power by looking at proportions that a test procedure rejects both null hypotheses or one particular hypothesis. We see that in all cases, CCOR test is better than SD test. CCOR test have similar power as the SU test in claiming

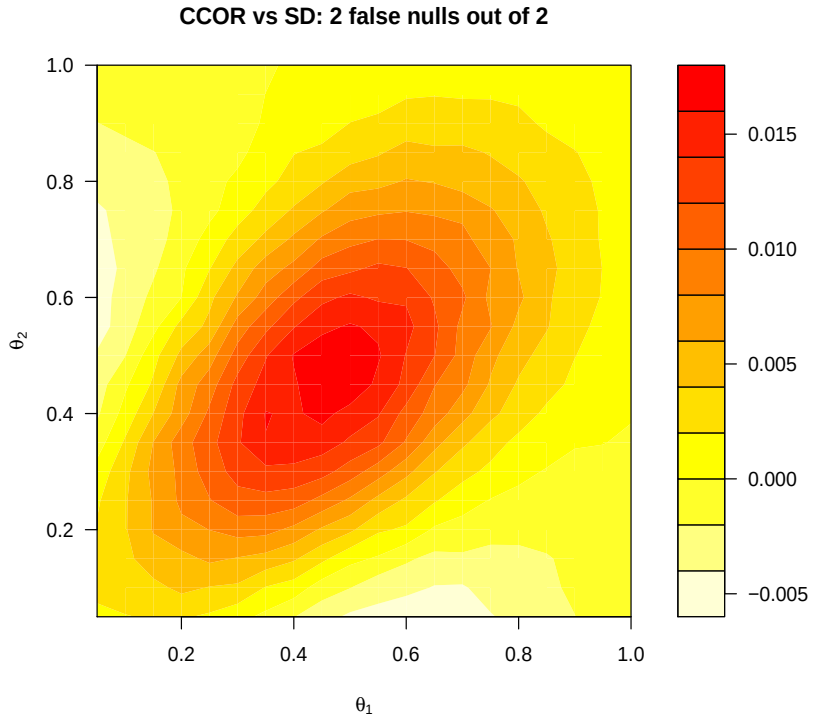

Figure 3.6: The power comparison among CCOR test and SD test in various (θ_1, θ_2) combinations from 10^5 simulations.

significance for both null hypotheses, but it has higher power in identifying significance for at least one hypothesis. The result confirms the conclusion that the CCOR test works well in all possible parameter configurations.

In a second simulation study, we compare $k = 4$ treatments with a control, with sample size 30 for each treatment and 40 for control. The performance of CCOR test, SD and SU tests are reported in Table 3.4 for selected $\boldsymbol{\theta}$ combinations. The proportion of rejecting each hypothesis and simultaneously rejecting l of the 4 hypotheses are compared ($l = 1, \ldots, 4$). The last column 'Avg' gives the average number of correct rejections. All results are based on $s = 10^6$ simulations. CCOR test is most efficient when all hypotheses are false as in the case $\boldsymbol{\theta} = (0.7, 0.7, 0.7, 0.7)$. CCOR test has the highest power for each individual null hypothesis.

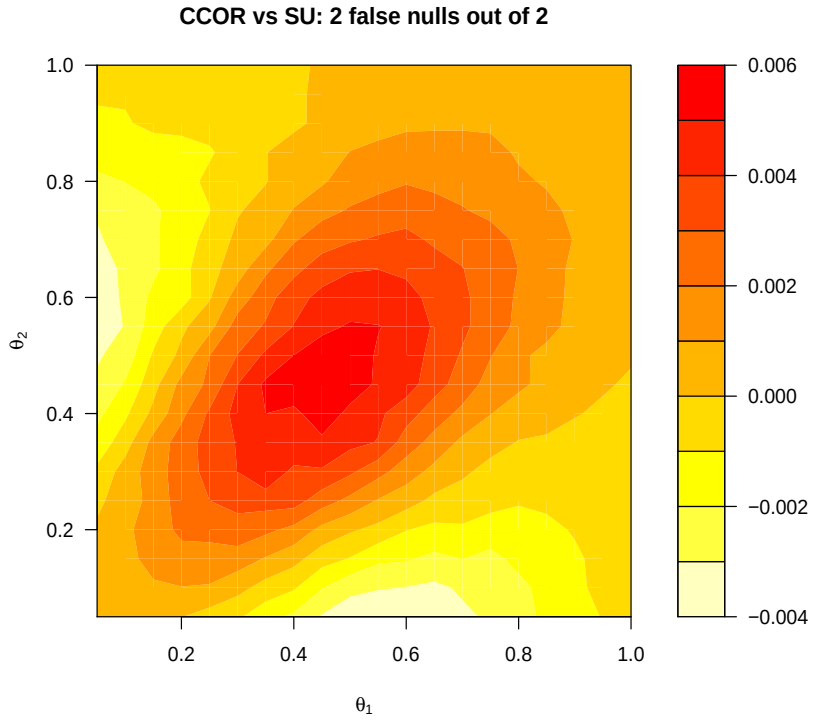

Figure 3.7: The power comparison among CCOR test and SU test in various (θ_1, θ_2) combinations from 10^5 simulations.

Its power to reject all four hypotheses is similar to SU test and higher than SD test in column $4/4$. In column $3/4$, CCOR has a power 1.4% higher than the other two tests. The average number of correct rejections of CCOR test is larger than SD and SU test by about 0.09 and 0.04, respectively. If most hypotheses are false, CCOR test also has better performance than SD/SU test as in the case of $\boldsymbol{\theta} = (0.55, 0.6, 0.65, 0)$. CCOR test has the highest power for each individual hypothesis. Although the power of rejecting all three false null hypotheses are slightly lower than SD/SU test (column $3/4$), its overall power catches up with an outstanding advantage of $2.1 \sim 2.4\%$ in column $2/4$. Thus the average number of correct rejections of CCOR test is higher than both SD and SU tests by about 0.02. If most hypotheses are true, CCOR test might fail to outperform SD/SU test as in the cases of $\boldsymbol{\theta} = (1, 1.2, 0, 0)$ and $\boldsymbol{\theta} = (0.7, 0, 0, 0)$. This is expected as indicated by the comparison of their rejection regions in Figure 3.5. However,

	$\theta_1 = 0.5, \theta_2 = 0.5$			$\theta_1 = 0.7, \theta_2 = 0.8$			$\theta_1 = 0.4, \theta_2 = 0.5$		
	both	only H_1	only H_2	both	only H_1	only H_2	both	only H_1	only H_2
CCOR	0.4972	0.1253	0.1250	0.8588	0.0271	0.0826	0.3948	0.0746	0.2052
SD	0.4838	0.1218	0.1214	0.8546	0.0259	0.0814	0.3819	0.0710	0.2032
SU	0.4974	0.1196	0.1193	0.8589	0.0256	0.0808	0.3950	0.0693	0.2000

Table 3.3: Performance of CCOR, SD and SU tests evaluated at selected $\boldsymbol{\theta}$ combinations from $s = 10^6$ simulations, $\alpha = 0.05$.

the advantages of SD/SU tests in these cases are not as significant as that of CCOR test in the previous two cases.

To conclude, we see from the rejection region plot in Figure 3.5 that none of the CCOR, SD and SU tests is uniformly more powerful than the others. Nevertheless, CCOR test has the optimal performance for arbitrary $\boldsymbol{\theta}$ configurations. It is particularly more powerful than the SD and SU tests when most hypotheses are false. Most importantly, CCOR test does not rely on the assumption of equal correlations to have strong control over the FWER as the SU test. As we will discuss in Section 3.6.2, CCOR test requires similar moderate computation as SU test. Whenever SU or SD test is affordable to be used, we recommend CCOR test as a better alternative. If computing effort is of major concern and conventional tests such as Holm's step-down test, Hochberg's step-up test or Hommel's test is used, one should proceed with caution since both Hochberg and Hommel procedures rest on the assumption of independent test statistics, which does not hold in the comparison with a control problem. Although Holm's procedure does not rely on such assumption, it is considered the most conservative one amongst all.

	$\boldsymbol{\theta} = (0.7, 0.7, 0.7, 0.7)$									$\boldsymbol{\theta} = (0.7, 0, 0, 0)$					
	H_1	H_2	H_3	H_4	$4/4$	$3/4$	$2/4$	$1/4$	Avg	H_1	H_2	H_3	H_4	$1/4$	Avg
CCOR	0.8648	0.8642	0.8649	0.8646	0.7132	0.1370	0.0741	0.0465	3.4585	0.7227	0.0212	0.0212	0.0211	0.7227	0.7227
SD	0.8428	0.8421	0.8426	0.8424	0.6932	0.1225	0.0842	0.0615	3.3700	0.7575	0.0206	0.0205	0.0205	0.7575	0.7575
SU	0.8562	0.8555	0.8559	0.8558	0.7135	0.1230	0.0732	0.0540	3.4235	0.7565	0.0214	0.0213	0.0212	0.7565	0.7565

	$\boldsymbol{\theta} = (0.55, 0.6, 0.65, 0)$								$\boldsymbol{\theta} = (1, 1.2, 0, 0)$						
	H_1	H_2	H_3	H_4	$3/4$	$2/4$	$1/4$	Avg	H_1	H_2	H_3	H_4	$2/4$	$1/4$	Avg
CCOR	0.6020	0.6707	0.7343	0.0481	0.4395	0.2562	0.1762	2.0070	0.9747	0.9972	0.0304	0.0302	0.9726	0.0266	1.9719
SD	0.5958	0.6646	0.7282	0.0454	0.4473	0.2319	0.1828	1.9886	0.9792	0.9977	0.0295	0.0293	0.9776	0.0217	1.9769
SU	0.5990	0.6676	0.7313	0.0475	0.4494	0.2351	0.1795	1.9979	0.9789	0.9977	0.0303	0.0301	0.9773	0.0220	1.9766

Table 3.4: Performance of CCOR, SD and SU tests in the case of 4 treatments, evaluated at selected $\boldsymbol{\theta}$ combinations from $s = 10^6$ simulations, $\alpha = 0.05$.

Column $l/4$ gives the rate of successfully rejecting l false null hypotheses. Column marked 'Avg' gives the average number of correct rejections.

3.5 Dose-response examples

To illustrate that the consonant closed likelihood ratio test, especially the CCOR test, is a powerful tool in dose-response studies, we first consider the data in Table 3.5 taken from Ruberg (1995). This pre-clinical dose-response study compares nine active doses with a zero-dose control. Six animals were investigated for each of the ten dose groups.

Group	Dosage	Sample	Mean	SD		mCCOR			CCOR			SD/SU		
	(mg/kg)	size	response	response	($\Delta =$)	6	5	0	6	5	0	6	5	0
0	0.0	6	25.5	2.6										
1	0.5	6	23.9	4.0										
2	1.0	6	27.7	3.3										
3	1.5	6	33.4	2.3				✓						
4	2.0	6	40.5	10.5		✓	✓	✓			✓		✓	✓
5	2.5	6	57.9	9.9		✓	✓	✓	✓	✓	✓	✓	✓	✓
6	3.0	6	74.4	14.6		✓	✓	✓	✓	✓	✓	✓	✓	✓
7	3.5	6	73.4	7.6		✓	✓	✓	✓	✓	✓	✓	✓	✓
8	4.0	6	73.5	4.5		✓	✓	✓	✓	✓	✓	✓	✓	✓
9	4.5	6	76.2	7.9		✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 3.5: Sample dose-response data from Ruberg (1995) and performance of closed test procedures for selected Δ values. Significant treatment effects are marked with "✓".

Usually in dose-response studies, people are interested in finding the effective doses which have mean responses greater than the control by a predefined value Δ. Since a location shift will not affect the sample variance, we can simply add Δ to all the responses in the control group and then apply the closed test procedures. In particular, the hypotheses of interests are

$$H_i : \theta_i \leqslant \Delta \quad \text{vs.} \quad A_i : \theta_i > \Delta; \quad i = 1, 2, \ldots, k. \tag{3.4}$$

When Δ is set to be 5, SD and SU tests both identify doses $4, \ldots, 9$ are effective while the CCOR test can only detect doses $5, \ldots, 9$. This result is caused by the fact that CCOR test is sensitive to the violation of equal variance assumption. As showed in Table 3.5, the sample standard deviations range from 2.3 of group 3 to 14.6 of group 6.

One possible modification to the CCOR test is to assume unequal variance and use individual sample variance as estimates. As a result, the decision on a particular intersection hypothesis will not be affected by the variance of unrelated treatment groups. This feature makes our test more robust to the heterogeneity of variance while still control the FWER at α. Specifically, the modified CCOR test (mCCOR) statistics derived from the likelihood ratio is defined as

$$S'_I = \frac{\prod_{i \in I \cup \{0\}} [(n_i - 1)s_i^2 + n_i(\bar{Y}_i - \tilde{Y}_i^I)^2]^{n_i}}{\prod_{i \in I \cup \{0\}} [(n_i - 1)s_i^2]^{n_i}},$$

where s_i^2 is the sample variance estimate of the ith group with $n_i - 1$ degrees of freedom. In Ruberg's data, the mCCOR test outperforms the original CCOR test and SD/SU tests for selected Δ values as showed in Table 3.5.

Hsu and Berger (1999) introduced a step-down method with pre-determined steps which takes advantage of suspected monotonicity in the response. Their procedure makes inferences in the form of "all doses greater than dose i are effective" and works well on Ruberg's data. More discussion and examples on this procedure can also be found in Bretz et al. (2008). However, this procedure is not designed to detect effective doses in the middle of an inverted-U (umbrella) response as presented in the following clinical trial example.

Dmitrienko et al. (2007a) provides an example of a Phase II clinical trial in 108 patients with mild to moderate asthma. The trial was conducted to compare three doses of an experimental drug to placebo. The drug was administered daily at doses 50, 250 and 1000 mg/day. The efficacy of the experimental drug was studied using spirometry measurements.

The primary efficacy endpoint was the forced expiratory volume in one second (FEV1). Table 3.6 displays the summary statistics of the experiment result and the graph shows the relationship between the mean FEV1 improvement and daily dose administered. The observation suggests an inverted-U shape dose-response relationship.

Group	Dose (mg/day)	Sample size	Mean response	SD response
0	0	28	0.049	0.37
1	50	27	0.244	0.43
2	250	26	0.282	0.43
3	1000	27	0.195	0.42

Table 3.6: FEV1 improvement in asthma trial example

Since this is not an equal sample size example, there is no analytical proof that SU procedure controls the familywise error rate strongly and the calculation of its critical constants is complicated. For the sake of simplicity, we use Hochberg's test as a replacement since it is considered as a Bonferroni approximation to SU test and provides similar power. For $\Delta = 0$ and $\alpha = 0.05$, only the CCOR test can identify the positive efficacy of the $250mg$ daily dosage. Both SD test and Hochberg's test result in retaining all hypotheses at $\alpha = 0.05$ and $\Delta = 0$.

3.6 Discussion

3.6.1 Consonant closed F test in mixed model with repeated measurements

Since most dose-response clinical trial nowadays have longitudinal data, we extend our consonant closed likelihood ratio test to linear mixed models with repeated measurements.

Consider a split-plot clinical trial design, where subjects are randomized to different treatment groups and followed across time. With $i = 0, 1, \ldots, k$ treatment groups ($i = 0$ as the control), $j = 1, \ldots, n$ subjects in treatment i and $t = 1, \ldots, T$ timepoints, we can write the model as

$$y_{ijt} = \mu + \tau_i + \beta_{j(i)} + \gamma_t + (\tau\gamma)_{it} + \varepsilon_{(ijt)},$$

where

$\mu =$ the overall mean;

$\tau_i =$ effect of treatment i ($\sum_i \tau_i = 0$);

$\beta_{j(i)} =$ individual difference component for subject j nested in treatment i, $\sim N(0, \sigma^2_\beta)$;

$\gamma_t =$ effect of time t ($\sum_t \gamma_t = 0$);

$(\tau\gamma)_{it} =$ interaction of treatment i and time t ($\sum_i (\tau\gamma)_{it} = \sum_t (\tau\gamma)_{it} = 0$);

$\varepsilon_{(ijt)} =$ random error for subject j in treatment i at time t, $\sim N(0, \sigma^2)$.

If the null hypothesis of no treatment by time interaction is rejected, then treatment and time effects are confounded with the interaction and cannot be tested separately. Otherwise, we can test the existence of treatment effect by $H : \tau_0 = \tau_1 = \ldots = \tau_k = 0$ using

$$F = \frac{\sum_{i=0}^{k} n(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2/k}{\sum_{i=0}^{k}\sum_{j=1}^{n}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot})^2/(k+1)(n-1)} \sim F_{k,(k+1)(n-1)}.$$

However, knowing the existence of treatment effect may not be enough, sometimes we are also interested in identifying which treatment group is different from the control. As a result, the k contrast tests of $H_i : \tau_i = \tau_0 = 0$ ($i = 1, \ldots, k$) should be performed. The consonant closed likelihood ratio test procedure can be applied in this context and the intersection hypotheses $H_I : \tau_i = \tau_0 = 0$, for all $i \in I \subset K = \{1, \ldots, k\}$ can be tested by the F type statistics

$$F_I = \frac{\sum_{i \in I \cup \{0\}} n(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2 \times \frac{k+1}{k|I|}}{\sum_{i=0}^{k}\sum_{j=1}^{n}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot})^2/(k+1)(n-1)}.$$

The consonance adjustment in this case is to reject H_I if $F_I > f_I$ and at least one of the hypotheses H_J, where $J \subset I$ and $|J| = |I| - 1$, is rejected. Then the critical constants f_I can be computed recursively using Monte-Carlo method similar to Section 3.3. The closed test procedure rejects a specific H_i ($i = 1, \ldots, k$) if all H_I with $I \ni i$ are rejected. The consonant closed likelihood ratio test guarantees that if a significant treatment effect exists, at least one of the treatments will be identified as significantly different from the control.

As a common drawback of fitting mixed model, the test procedure in this case is inexact. The reasons of the inexactness is partly because that we need to treat the estimated variances-covariance as if they were fixed and known in the computation of critical constants of consonant closed likelihood ratio test (when generating simulation samples). However, the logic of our consonant closed test does not depends on the structure of the variance-covariance matrix, which requires that the rejection of an intersection hypothesis be accompanied by at least one rejection among its component hypotheses. As a result, as long as there exists a reasonable estimate for the variance-covariance matrix, our consonant closed likelihood ratio test can be applied to any general mixed model. For more details about multiple comparison in mixed model, please refer to Westfall et al. (1999); Westfall and Young (1993) and the reference therein.

Also note that in Phase II clinical trials a strict type I error rate control of dose-response study is often not mandatory and modeling approaches are reasonable alternatives. Bretz et al. (2008) gave a review of modeling approaches and hybrid approaches combining multiple comparisons and modeling techniques. For interested readers, we refer to Bretz et al. (2008, 2005) and the references therein for details.

3.6.2 Computational Issue

The critical constants and p-values of consonant closed likelihood ratio test are computed using Monte-Carlo method. The accuracy of such computation depends on the number of Monte-Carlo simulation m and the number of treatments k. In general, the critical constant approach is recommended over the p-value approach. Since the critical constants of CCF and CCOR tests are computed recursively as in (3.2) and (3.3), the p-value computation of intersection hypothesis H_I relies on the knowledge of all critical constants in lower level intersection hypotheses H_J ($J \subset I$). In other words, to compute the p-values for each intersection hypothesis, the critical constants need to be calculated first.

Suppose m sets of Monte-Carlo samples are generated, we can calculated all the likelihood ratio test statistics (F or S, for CCF and CCOR test respectively) for each set of the sample. The critical constant is then the estimated quantile of the underlying distribution. Several algorithms are reviewed and compared in Hyndman and Fan (1996). We applied the default method in R and S-PLUS. Usually a quantile estimator has a slow convergence rate $\mathcal{O}(n^{-1/2})$. To evaluate the accuracy of our critical constant computation, a bootstrap variance estimation is used. From this m Monte-Carlo samples, bootstrap replicates of size m are sampled with replacement. The sample quantile is estimated again using each bootstrap sample. A total of $b = 1000$ bootstrap samples are collected and the variance is computed.

In computing the critical values used in examples of Section 3.3 and 3.4, $m = 10^7$ is used. The bootstrap standard deviation of those estimates are in the neighborhood of 0.001. Table 3.1 and 3.2 is constructed with $m = 10^6$, the bootstrap standard deviation of those critical constants are less than 0.01. All computations are performed on the high-performance computing infrastructure at the UCR Institute for Integrative Genome Biology. On a 2.5GHz core, the computation of CCOR test critical constants takes 225 seconds and 1183 seconds for $k = 2$ and

$k = 4$ treatments with $m = 10^6$. We write our own R program for SU test critical constants following algorithm in Kwong and Liu (2000) and it takes 149 seconds and 1020 seconds for $k = 2$ and $k = 4$ treatments with $m = 10^5$. We believe the computing effort of CCOR test is moderate and at least similar to (if not less than) the SU test.

Chapter 4

A New Partition Testing Strategy for Multiple Primary and Secondary Endpoints in Regulatory Applications

4.1 Introduction to multiple endpoints problem in clinical trials

Many clinical trials involve multiple endpoints and dose-to-control comparisons. The endpoints can be generally classified as primary and secondary. A primary endpoint describes how the most important aspects of the disease are affected by the intervention. An intervention's effect on the secondary endpoint can further help prescribing physicians in identifying suitable treatments for their patients (Offen et al., 2007). However, a significant improvement in a secondary endpoint alone is not considered as substantial evidence of clinical benefit. Based on this understanding, multiple hypotheses testing strategies are proposed following the common

rules that,

1. the efficacy in the secondary endpoints is considered only if efficacy in the primary endpoints has been shown;

2. inference made in primary endpoints is not affected by the inference made in secondary endpoints. (4.1)

Hung and Wang (2009) suggest that these rules should be enforced only within each given dose level. Specifically, they illustrate the idea through an example: If there is sufficient evidence to conclude that the high dose is effective with respect to the primary endpoint, then the secondary endpoint testing for the high dose should not be subjected to proving that the low dose is also effective with respect to the primary endpoint. These rules play a key role in developing testing procedures for multiple endpoints. Several extensions of the rules exist when the structure of multiple endpoints becomes more complex.

Currently, the sequential gatekeeping procedure (Dmitrienko et al., 2003; Chen et al., 2005; Dmitrienko et al., 2006b, 2007b) is widely used in clinical trials with multiple endpoints. This procedure is based on the closed testing principle (Gabriel, 1969) which tests all intersection hypotheses in the family and rejects a hypothesis only if all intersection hypotheses implying it are rejected. For each intersection hypothesis, the weighted Bonferroni test is applied and the two rules are implemented by assigning different weight to each component of the intersection. Several variations of the gatekeeping procedure exist, such as serial, parallel and tree-structured gatekeepings. They are characterized by the different weight selection algorithms. The two major drawbacks of gatekeeping method are the complexity of weight selection algorithm and the limited power due to Bonferroni approximation.

In the direction to simplify the steps in gatekeeping procedure, Bretz et al. (2009) recently proposed a graphical approach for the sequentially rejective multiple test procedures

in which gatekeeping procedures are special cases. Their approach are represented by directed, weighted graphs where each node corresponds to an elementary hypothesis that a certain dose-endpoint combination lacks efficacy. This approach provides shortcut to the Bonferroni-type closed test, thus making extensive computer programming unnecessary. It is also a great tool to communicate with clinical team. However, this procedure is limited to Bonferroni tests only.

Yet another variation of the gatekeeping method tests each intersection hypothesis through the single-step Dunnett test instead of Bonferroni test (Dmitrienko et al., 2006a). This procedure assumes that the joint distribution of the test statistics from different endpoints can be approximated by a multivariate t-distribution and the correlation between endpoints can be well estimated. It provides a power advantage over the Bonferroni gatekeeping procedure. However, the assumptions made by this procedure are liberal, which will result in the lost of strong control over the familywise error rate (FWER). A Dunnett-Bonferroni-based gatekeeping procedure is proposed as an improvement (Xu et al., 2009) which applies Dunnett test within endpoint and Bonferroni inequality across endpoints. This procedure does not rely on the unrealistic assumptions which might be challenged by regulatory agencies but provides almost the same performance.

More recently, Liu and Hsu (2009) proposed a partition testing approach which reduces the number of intersection hypotheses by partitioning the null space for each dose. They follow the rules in (4.1) within each given dose level, which they referred to as Decision Path. The resulting procedure tests $(m+1)^k - 1$ intersection hypotheses in clinical study with m hierarchically ordered endpoints and k doses. The number of intersection hypotheses tested in a typical gatekeeping procedure for the same experiment is $2^{mk} - 1$. The partition testing formulates fewer hypotheses to start with. For each intersection hypothesis, the joint distribution within endpoint is used along with independence assumption or Bonferroni inequality across endpoints.

Despite the different appearance of the above procedures, they all share the common union-intersection setting, in which an intersection hypothesis is rejected if at least one of the component hypotheses is rejected. Specifically, the Bonferroni-based gatekeeping procedure rejects an intersection hypothesis if at least one of the component hypotheses is rejected by the Bonferroni test with specified weight. The graphical approach admits a shortcut, which is also derived from the underlying weighted Bonferroni test for each intersection hypothesis. The Dunnett-based gatekeeping procedure and the partition testing rejects an intersection hypothesis if at least one of the component t statistics are significantly large.

In this chapter, We propose a new partition testing approach which focuses on testing each intersection hypothesis with likelihood ratio test. As discussed in Chapter 3, when using likelihood ratio test, the logical consistency of consonance may be lost. We propose a consonance adjustment technique to retain the consistency in analog to the UI method. The resulting new partition testing procedure provides power advantage over gatekeeping procedure (GK) and the original decision path partition testing (DP).

This chapter is organized as follows. Section 4.2 reviews the partitioning decision paths proposed by Liu and Hsu (2009) and extend it to alternative primary endpoints and co-primary endpoints problems. Section 4.3 introduces the likelihood ratio test and the consonance adjustment. The adjustment procedure is illustrated there with a simple example. Section 4.4 demonstrates the new procedure in two advanced cases, with simulation study comparing to existing methods. Section 4.5 draw some concluding remarks.

4.2 The Partitioning Decision Path Principle

4.2.1 Hierarchical endpoints

Suppose k doses of an experiment drug are compared to a control on a single primary endpoint and a single secondary endpoint. Let μ_i^E denote the mean response of group i for endpoint E, $i = 0, 1, \cdots, k$, $E = P, S$ (for primary and secondary endpoint), where $i = 0$ denotes the control group. Define the difference in mean efficacy measurement between each dose group i and the control group for endpoint E as $\theta_i^E = \mu_i^E - \mu_0^E$, $E = P, S$. Then the original elementary hypotheses of interest are

$$\begin{aligned} H_i^P : \theta_i^P \leq \delta^P \quad \text{vs} \quad A_i^P : \theta_i^P > \delta^P \\ H_i^S : \theta_i^S \leq \delta^S \quad \text{vs} \quad A_i^S : \theta_i^S > \delta^S \end{aligned} \tag{4.2}$$

for any $i = 1, \cdots, k$, if a larger response indicates better treatment. To be general, θ_i^E is compared to a clinically meaningful difference δ^E for endpoint E. Without confusion, we use capital letter H (or A) to denote both a hypothesis and its corresponding parameter space.

A sequence of potential inferences is termed as a *path* in Liu and Hsu (2009). In the hierarchical endpoints problem, such path does not make inference for lower ordered endpoint before efficacy is established for all the higher ordered endpoints FOR A GIVEN DOSE. Given the decision paths for each dose, the parameter space is partitioned in two stages, namely, path partition and disjointness partition Liu and Hsu (2009).

Path partition is within each dose i, starting with the primary endpoint null hypothesis $H_i^P : \theta_i^P \leq \delta^P$. If H_i^P is rejected, then efficacy in the primary endpoint is established for dose i. Following the path to the secondary endpoint, $H_i^{\star S} = H_i^S \backslash H_i^P : \theta_i^S \leq \delta^S$ and $\theta_i^P > \delta^P$ is tested instead of H_i^S. If both H_i^P and $H_i^{\star S}$ are rejected, then efficacy in both primary and secondary endpoints are established. However, if H_i^P is retained, then even if $H_i^{\star S}$ is rejected, no inference should be made. Since H_i^P and $H_i^{\star S}$ are disjoint, the FWER is strongly controlled

if each is tested at level α. The path partition can be visualized in Figure 4.1. Besides H_i^P and $H_i^{\star S}$, the remaining part of parameter space is $\theta_i^P > \delta^P$ and $\theta_i^S > \delta^S$, which does not need to be tested and is denoted by H_i^A.

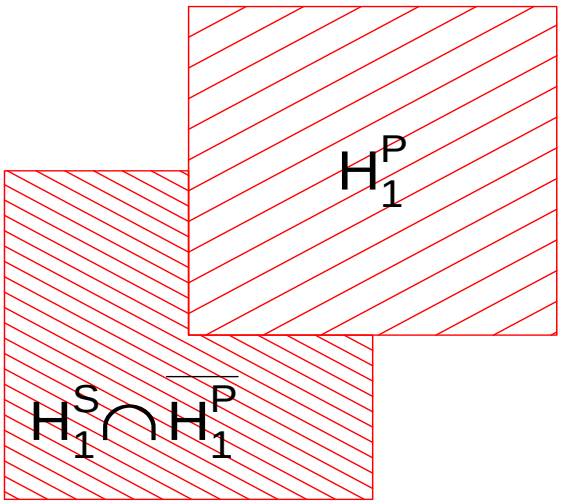

Figure 4.1: The path partition of the partition testing in the case of one primary endpoint and one secondary endpoint.

Between paths, the partitioned hypotheses are not disjoint and their intersections are considered. This is done through the disjointness partition. In the simple case with $k = 2$, the family of hypotheses for dose i is $\mathscr{H}_i = \{H_i^P, H_i^{\star S}, H_i^A\}$. The intersections between the two families $\mathscr{H}_1$ and $\mathscr{H}_2$ (for 2 doses) partition the parameter space into 9 parts, with the intersection $H_1^A \cap H_2^A$ need not be tested. The result of disjointness partition can be visualized in Figure 4.2. Since each intersection concerns selecting one partition from each of the two families $\mathscr{H}_1$

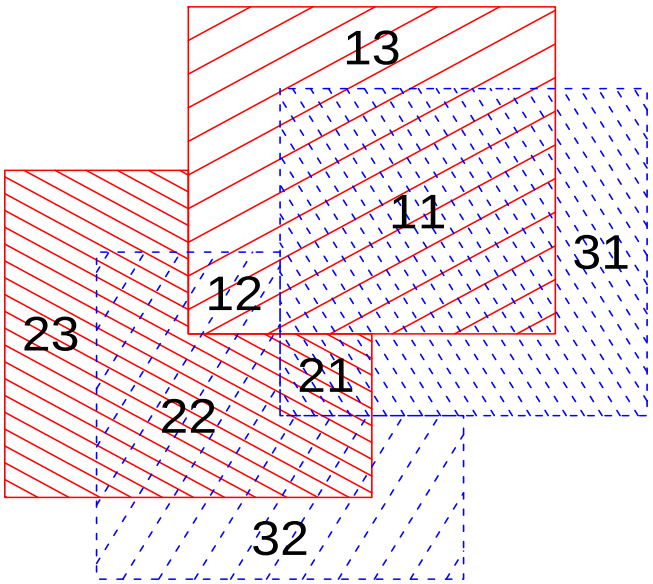

Figure 4.2: The disjointness partition of the partition testing in the case of two doses with one primary endpoint and one secondary endpoint. The labeling of disjointness partition follows that described in (4.3).

and $\mathscr{H}_2$, they can be labeled conveniently as $\Theta_{e_1e_2} = \cap_{i=1}^{2}\Theta_i^{e_i}$ where

$$\Theta_i^{e_i} = \begin{cases} \Theta_i^P = \{\theta_i^P \leq \delta^P\} & \text{if } e_i = 1 \\ \Theta_i^{\star S} = \{\theta_i^P > \delta^P \text{ and } \theta_i^S \leq \delta^S\} & \text{if } e_i = 2\,, \\ \Theta_i^A = \{\theta_i^P > \delta^P \text{ and } \theta_i^S > \delta^S\} & \text{if } e_i = 3 \end{cases} \qquad i = 1, 2. \tag{4.3}$$

Thus the notation $\Theta_{e_1e_2}$ represents the lack of efficacy for certain dose-endpoint combinations. The sub-indices can be extended to $e_1, \ldots, e_k$ in the case of $k > 2$ doses. Proved as Theorem 1 in Liu and Hsu (2009), for a clinical trial with m (hierarchically) ordered endpoints and k (unordered) doses, partition testing following the Decision Path Principle leads to testing $(m +$

$1)^k - 1$ hypotheses that partition the parameter space. The final inference is to reject H_i^P if all of the partition hypotheses $\Theta_{e_1 e_2}$ with $e_i = 1$ are rejected. In addition, if all the partition hypotheses $\Theta_{e_1 e_2}$ with $e_i = 2$ are rejected, then we can reject H_i^S as well. The inference automatically respects the prespecified paths and controls FWER strongly at level α, providing each partition hypothesis is tested at level α.

4.2.2 Alternative primary endpoints

When a treatment is deemed efficacious if it improves at least one of the multiple primary endpoints, each primary endpoint is an alternative to other primary endpoints in determining the efficacy of the treatment. It is thus called *alternative primary endpoints*.

For simplicity, we focus on the case with two alternative primary endpoints and one secondary endpoint. Denote the primary endpoints as P1 and P2, the secondary endpoint as S. Then the hypotheses in (4.4) are the original elementary null hypotheses,

$$\begin{aligned} H_i^{P1} &: \theta_i^{P1} \leq \delta^{P1} \\ H_i^{P2} &: \theta_i^{P2} \leq \delta^{P2}, \quad i = 1, 2, \cdots, k \\ H_i^{S} &: \theta_i^{S} \leq \delta^{S} \end{aligned} \tag{4.4}$$

For each dose i, the decision path of Liu and Hsu (2009) suggests to consider the hypotheses H_i^{P1}, H_i^{P2} and $H_i^{\star S} = H_i^S \backslash (H_i^{P1} \cap H_i^{P2})$ in turn. The efficacy in primary endpoint P1 (P2) can be claimed if H_i^{P1} (H_i^{P2}) is rejected. The efficacy in the secondary endpoint is established if $H_i^{\star S}$ and at least one of H_i^{P1} and H_i^{P2} are rejected. It guarantees that the secondary endpoint efficacy can be claimed only if at least one of the primary endpoint has been proved efficacious. However, the three hypotheses H_i^{P1}, H_i^{P2} and $H_i^{\star S}$ are not disjoint in this case, thus not a partition of the parameter space of dose i. To extend the decision path principle, we consider a natural partition based on the three hypotheses. First consider the $\cap$-closed family of hypotheses for dose i, $\mathscr{H}_i = \{H_i^{P1}, H_i^{P2}, H_i^{P1} \cap H_i^{P2}, H_i^{\star S}, H_i^{P1} \cap$

$H_i^{\star S}, H_i^{P2} \cap H_i^{\star S}, H_i^A\}$, where $H_i^A = \{\theta_i^{P1} > \delta^{P1}$ and $\theta_i^{P2} > \delta^{P2}$ and $\theta_i^S > \delta^S\}$. Then consider the natural partition of $\mathscr{H}_i$ as defined in Finner and Strassburger (2002), denoted as $\Theta_i = \{\Theta_i^{P1}, \Theta_i^{P2}, \Theta_i^{P1P2}, \Theta_i^S, \Theta_i^{P1S}, \Theta_i^{P2S}, \Theta_i^A\}$, where

$$\begin{aligned}
\Theta_i^{P1} &= \{\theta_i^{P1} \leq \delta^{P1} \text{ and } \theta_i^{P2} > \delta^{P2} \text{ and } \theta_i^S > \delta^S\},\\
\Theta_i^{P2} &= \{\theta_i^{P1} > \delta^{P1} \text{ and } \theta_i^{P2} \leq \delta^{P2} \text{ and } \theta_i^S > \delta^S\},\\
\Theta_i^{P1P2} &= \{\theta_i^{P1} \leq \delta^{P1} \text{ and } \theta_i^{P2} \leq \delta^{P2}\},\\
\Theta_i^S &= \{\theta_i^{P1} > \delta^{P1} \text{ and } \theta_i^{P2} > \delta^{P2} \text{ and } \theta_i^S \leq \delta^S\},\\
\Theta_i^{P1S} &= \{\theta_i^{P1} \leq \delta^{P1} \text{ and } \theta_i^{P2} > \delta^{P2} \text{ and } \theta_i^S \leq \delta^S\},\\
\Theta_i^{P2S} &= \{\theta_i^{P1} > \delta^{P1} \text{ and } \theta_i^{P2} \leq \delta^{P2} \text{ and } \theta_i^S \leq \delta^S\},\\
\Theta_i^A &= \{\theta_i^{P1} > \delta^{P1} \text{ and } \theta_i^{P2} > \delta^{P2} \text{ and } \theta_i^S > \delta^S\}.
\end{aligned} \tag{4.5}$$

Θ_i partitions the parameter space of dose i into 7 disjoint pieces.

Then apply the disjointness partition, the entire parameter space is partitioned into 7^k pieces for k doses, with the one $\cap_{i=1}^k \Theta_i^A$ need not be tested. Notice the number of intersection hypotheses to be tested is $7^k - 1$, which is still less than the amount in gatekeeping procedure ($2^{3k} - 1 = 8^k - 1$). Consider the hypotheses of interest, H_i^{P1}, H_i^{P2} and H_i^S. They can be expressed as unions of certain partition hypotheses in (4.5). The final inference is based on collating the result testing each partition hypothesis.

4.2.3 Co-primary endpoints

There are many disorders for which the regulatory agencies have required a treatment to demonstrate significant effect on all multiple endpoints, before accepting the treatment's clinical benefit. This is usually referred to as a *co-primary endpoints* problem or reverse multiplicity problem, which according to the intersection-union (IU) test does not need multiplicity adjustment of any kind. Such requirement could lead to substantial reduction in the study's power

to conclude the efficacy of a treatment. Several publications discussed this issue and proposed possible solutions. To name a few, Offen et al. (2007); Hung and Wang (2009); Chuang-Stein et al. (2007) and Eaton and Muirhead (2007).

We extend the partition testing to the case with two co-primary endpoints and a secondary endpoint. The hypothesis H_i^S in the secondary endpoint will not be tested unless both H_i^{P1} and H_i^{P2} are rejected (denoted the co-primary endpoints as P1 and P2). The hypotheses are defined the same way as in (4.4). The decision path considers the hypotheses H_i^{P1}, H_i^{P2} and $H_i^{\star S} = H_i^S \backslash (H_i^{P1} \cup H_i^{P2})$ in turn. Notice that $H_i^{\star S}$ defined in co-primary endpoint problems are different from that of alternative primary endpoint problems. The efficacy in primary endpoint P1 (P2) can be claimed if H_i^{P1} (H_i^{P2}) is rejected. The efficacy in the secondary endpoint is established if $H_i^{\star S}$, H_i^{P1} and H_i^{P2} are all rejected. Again, the three hypotheses are not disjoint. We construct the natural partition as follows. First determine the $\cap$-closed family of hypotheses for dose i, $\mathscr{H}_i = \{H_i^{P1}, H_i^{P2}, H_i^{P1} \cap H_i^{P2}, H_i^{\star S}, H_i^A\}$, where H_i^A is the hypothesis that all endpoints in dose i are efficacious. Then the natural partition is $\Theta_i = \{\Theta_i^{P1}, \Theta_i^{P2}, \Theta_i^{P1P2}, \Theta_i^{\star S}, \Theta_i^A\}$, where

$$
\begin{aligned}
\Theta_i^{P1} &= \{\theta_i^{P1} \leq \delta^{P1} \text{ and } \theta_i^{P2} > \delta^{P2}\},\\
\Theta_i^{P2} &= \{\theta_i^{P1} > \delta^{P1} \text{ and } \theta_i^{P2} \leq \delta^{P2}\},\\
\Theta_i^{P1P2} &= \{\theta_i^{P1} \leq \delta^{P1} \text{ and } \theta_i^{P2} \leq \delta^{P2}\},\\
\Theta_i^{S} &= \{\theta_i^{P1} > \delta^{P1} \text{ and } \theta_i^{P2} > \delta^{P2} \text{ and } \theta_i^{S} \leq \delta^{S}\},\\
\Theta_i^{A} &= \{\theta_i^{P1} > \delta^{P1} \text{ and } \theta_i^{P2} > \delta^{P2} \text{ and } \theta_i^{S} > \delta^{S}\}.
\end{aligned}
\tag{4.6}
$$

Θ_i partitions the parameter space of dose i into 5 disjoint pieces. Apply the disjointness partition, the entire parameter space is partitioned into 5^k pieces for k doses. Thus there are $5^k - 1$ intersection hypotheses to be tested. Compare the natural partitions in (4.6) and (4.5), the requirements of alternative primary endpoints and co-primary endpoints make a clear difference.

The final inference collates the results testing each partition hypothesis in (4.6).

In this section, We generalized the decision path principle of Liu and Hsu (2009) to the cases with alternative primary endpoints and co-primary endpoints. This extension is required mainly because the $\cap$-closed family of hypotheses $\mathscr{H}_i$ in those cases is not a partition of the parameter space of dose i. Thus a natural partition step is taken and Θ_i instead of $\mathscr{H}_i$ is used in the following disjointness partition step. For the sake of notation consistency, we will use Θ_i to denote the disjoint partitions of the parameter space of dose i in all cases. Thus in the case of hierarchically ordered multiple endpoints, $\mathscr{H}_i = \Theta_i$ are equivalent.

The corresponding techniques used in gatekeeping procedure for alternative and co-primary endpoints are referred to as parallel and serial gatekeeping, respectively. Section 4.2.2 and 4.2.3 are presented with two primary endpoints, for scenarios with more than two primary endpoints and combination with hierarchical endpoints, the decision path principle can be extended easily as we discussed.

4.3 Likelihood Ratio Test and Consonance Adjustment: Fundamentals

In this section, we focus our discussion on the scenario with a primary endpoint and a secondary endpoint for $k = 2$ doses. The testing procedure can be easily extended to more general scenarios. Some extensions are discussed in Section 4.4.

4.3.1 Likelihood Ratio Test

Suppose two hierarchically ordered endpoints (primary and secondary) are considered in a confirmatory clinical trial. Following the notations defined in Section 4.2.1 with $k = 2$ doses, the decision path partition results in $3^k - 1 = 8$ null hypotheses of interest in (4.3).

The partition testing of Liu and Hsu (2009) rejects an intersection hypothesis $\Theta_{e_1e_2}$ if at least one of the t statistics of the component hypotheses is significantly large. Instead of using the union-intersection setting, we propose to apply the likelihood ratio test for each intersection hypothesis. Throughout the years, likelihood ratio tests are regarded as computationally intractable (Tang et al., 1989; Wassmera et al., 1999). However, after decades of development in computer technology, the amount of computation required by likelihood ratio tests are no longer a hassle. Our consonant likelihood ratio test shows a significant power advantage in the center direction when most null hypotheses are false.

The likelihood ratio statistic is the ratio of likelihood functions maximized under the null hypothesis and the global hypothesis. For each partition $\Theta_{e_1e_2}$, the null and alternative hypotheses can be written as

$$H_0 : \Theta_{e_1e_2} \qquad \text{vs} \qquad H_a : \Omega \backslash \Theta_{e_1e_2},$$

where Ω is the global hypothesis with no constraint on any parameter.

Specifically, the 8 null hypotheses of interest are summarized in Table 4.1. The ideal situation that both high and low doses are efficacious in both endpoints, Θ_{33} does not need to be tested.

For each partition hypothesis, it is a common partition testing practice to test only the part of null hypothesis states $\theta_i^E \leq \delta^E$. For example, instead of testing the hypothesis Θ_{12} that states

$$\theta_1^P \leq \delta^P \text{ and } \theta_2^P > \delta^P \text{ and } \theta_2^S \leq \delta^S,$$

the hypothesis H_{12} that states

$$\theta_1^P \leq \delta^P \text{ and } \theta_2^S \leq \delta^S$$

is tested. A level-α test for H_{12} is also a level-α test for Θ_{12}. The null hypotheses actually

Hypotheses of interest	θ_1^P	θ_1^S	θ_2^P	θ_2^S	Hypotheses tested	θ_1^P	θ_1^S	θ_2^P	θ_2^S
Θ_{11}	$\leq \delta^P$	-	$\leq \delta^P$	-	H_{11}	$\leq \delta^P$	-	$\leq \delta^P$	-
Θ_{12}	$\leq \delta^P$	-	$> \delta^P$	$\leq \delta^S$	H_{12}	$\leq \delta^P$	-	-	$\leq \delta^S$
Θ_{21}	$> \delta^P$	$\leq \delta^S$	$\leq \delta^P$	-	H_{21}	-	$\leq \delta^S$	$\leq \delta^P$	-
Θ_{13}	$\leq \delta^P$	-	$> \delta^P$	$> \delta^S$	H_{13}	$\leq \delta^P$	-	-	-
Θ_{31}	$> \delta^P$	$> \delta^S$	$\leq \delta^P$	-	H_{31}	-	-	$\leq \delta^P$	-
Θ_{22}	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$\leq \delta^S$	H_{22}	-	$\leq \delta^S$	-	$\leq \delta^S$
Θ_{23}	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$> \delta^S$	H_{23}	-	$\leq \delta^S$	-	-
Θ_{32}	$> \delta^P$	$> \delta^S$	$> \delta^P$	$\leq \delta^S$	H_{32}	-	-	-	$\leq \delta^S$

Table 4.1: The null hypotheses of interest and null hypotheses actually tested featuring 2 doses and primary-secondary endpoints. A short dash '$-$' indicates no constraint on that parameter.

tested are also listed in Table 4.1.

Assume the samples from dose group i have the following bivariate normal distribution

$$\mathbf{Y}_{ij} = \begin{pmatrix} Y_{ij}^P \\ Y_{ij}^S \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left(\begin{pmatrix} \mu_i^P \\ \mu_i^S \end{pmatrix}, \begin{pmatrix} \sigma_P^2 & \rho\sigma_P\sigma_S \\ \rho\sigma_P\sigma_S & \sigma_S^2 \end{pmatrix} \right) \tag{4.7}$$

where $i = 0, 1, 2$ and $j = 1, \ldots, n_i$ are index for the n_i subjects in dose group i. The variance of primary and secondary endpoints, σ_P^2 and σ_S^2 are assumed unknown. In general, there are three approaches to deal with the unknown correlation coefficient ρ between the endpoints. (1) Find a consistent estimate of ρ and use it as if it is the true value; (2) Apply Bonferroni inequality which does not depend on ρ; (3) Assume independence ($\rho = 0$). Since ρ can not always be well estimated given the observed data, the second and third options are of more interest in practice. Bonferroni-based gatekeeping procedures apply the second option across endpoints

and dose levels, it is conservative and controls the FWER strongly. Liu et al. (2007) proved that assuming independence among the t statistics results in conservative testing procedures when the true correlation is positive. When there is no strong belief of positive association among endpoints, Liu and Hsu (2009) suggests using Bonferroni inequality in their partition testing procedure across endpoints. Since likelihood ratio test focus on ellipsoidal acceptance regions instead of rectangles, we are able to show that

Theorem 7 *Under the independence assumption* $(\rho = 0)$*, the likelihood ratio test is conservative and controls the FWER strongly.*

To prove Theorem 7, we need the following notations.

Denote the likelihood function of $\{\mathbf{Y}_{ij} : i = 0, 1, 2; j = 1, \ldots, n_i\}$ as $L(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\sigma})$. Use $\bar{Y}_i^P$, $\bar{Y}_i^S$, $\hat{\sigma}_P$ and $\hat{\sigma}_S$ to denote the MLEs of μ_i^P, μ_i^S, σ_P and σ_S under the global hypothesis Ω. Use $\tilde{Y}_i^P$, $\tilde{Y}_i^S$, $\tilde{\sigma}_P$ and $\tilde{\sigma}_S$ to denote the constraint MLEs under the null hypothesis $H_{e_1e_2}$. The distributions and computation techniques of constraint MLEs are extensively studied in Robertson et al. (1988). Then the test statistic for $H_{e_1e_2}$ is

$$\lambda_{e_1e_2} = \frac{\sup_{\boldsymbol{\theta} \in H_{e_1e_2}} L}{\sup_{\boldsymbol{\theta} \in \Omega} L} = \left(\frac{\tilde{\sigma}_P^2 \tilde{\sigma}_S^2}{\hat{\sigma}_P^2 \hat{\sigma}_S^2} \right)^{-\frac{\sum n_i}{2}} = \left(\frac{\sum_{ij} (y_{ij}^P - \tilde{y}_i^P)^2 \sum_{ij} (y_{ij}^S - \tilde{y}_i^S)^2}{\sum_{ij} (y_{ij}^P - \bar{y}_i^P)^2 \sum_{ij} (y_{ij}^S - \bar{y}_i^S)^2} \right)^{-\frac{\sum n_i}{2}}. \tag{4.8}$$

Proof of Theorem 7. Without lose of generality, we prove Theorem 7 in the context of primary-secondary endpoints with 2 doses. Use the notation in (4.8), to test $\Theta_{e_1e_2}$, it is sufficient to prove for any $\boldsymbol{\theta} \in \Theta_{e_1e_2}$,

$$P_{\boldsymbol{\theta},\rho} \left(\frac{\hat{\sigma}_P^2 \hat{\sigma}_S^2}{\tilde{\sigma}_P^2 \tilde{\sigma}_S^2} \geq c \right) \geq P_{\boldsymbol{\theta},\rho=0} \left(\frac{\hat{\sigma}_P^2 \hat{\sigma}_S^2}{\tilde{\sigma}_P^2 \tilde{\sigma}_S^2} \geq c \right).$$

Since the left hand side is equal to $\int P_{\boldsymbol{\theta},\rho}(\frac{\hat{\sigma}_P^2}{\tilde{\sigma}_P^2} \geq c/t, \frac{\hat{\sigma}_S^2}{\tilde{\sigma}_S^2} = t)dt$ and the right hand side is equal to $\int P_{\boldsymbol{\theta}}(\frac{\hat{\sigma}_P^2}{\tilde{\sigma}_P^2} \geq c/t) P_{\boldsymbol{\theta}}(\frac{\hat{\sigma}_S^2}{\tilde{\sigma}_S^2} = t)dt$. We only need to show the inequality of

$$P_{\boldsymbol{\theta},\rho} \left(\frac{\hat{\sigma}_P^2}{\tilde{\sigma}_P^2} \geq c/t, \frac{\hat{\sigma}_S^2}{\tilde{\sigma}_S^2} = t \right) \geq P_{\boldsymbol{\theta}} \left(\frac{\hat{\sigma}_P^2}{\tilde{\sigma}_P^2} \geq c/t \right) P_{\boldsymbol{\theta}} \left(\frac{\hat{\sigma}_S^2}{\tilde{\sigma}_S^2} = t \right). \tag{4.9}$$

This is a special case of the Gaussian correlation conjecture which states that for any two symmetric convex sets A and B in a separable Banach space E and for any centered Gaussian measure μ on E,

$$\mu(A \cap B) \geq \mu(A)\mu(B). \tag{4.10}$$

For a history of this conjecture, we refer to Das Gupta et al. (1972) and Li and Shao (2001). Schechtman et al. (1998) showed that the conjecture is true whenever the sets are symmetric and ellipsoid. Hargé (1998) proved that (4.10) holds if one set is symmetric ellipsoid and the other is simply symmetric convex. Notice that

$$\frac{\hat{\sigma}_P^2}{\tilde{\sigma}_P^2} = \frac{\sum_{ij}(y_{ij}^P - \bar{y}_i^P)^2}{\sum_{ij}(y_{ij}^P - \tilde{y}_i^P)^2}, \qquad \frac{\hat{\sigma}_S^2}{\tilde{\sigma}_S^2} = \frac{\sum_{ij}(y_{ij}^S - \bar{y}_i^S)^2}{\sum_{ij}(y_{ij}^S - \tilde{y}_i^S)^2}$$

are symmetric to $\bar{\mathbf{y}}^P$ and $\bar{\mathbf{y}}^S$ respectively, where $\bar{\mathbf{y}}^E = (\bar{y}_0^E, \ldots, \bar{y}_0^E, \ldots, \bar{y}_2^E, \ldots, \bar{y}_2^E)$ with the component $\bar{y}_i^E$ repeats itself for n_i times, $i = 0, 1, 2$ and $E = P, S$. This can be proved by the fact that the constraint MLEs $\bar{y}_i^P$, $\tilde{y}_i^P$ ($\bar{y}_i^S$, $\tilde{y}_i^S$) are fixed for each set of $(\bar{y}_0^P, \bar{y}_1^P, \bar{y}_2^P)$ and $(\bar{y}_0^S, \bar{y}_1^S, \bar{y}_2^S)$ values. In other words, for various values $\{y_{ij}^E : i = 0, 1, 2; j = 1, \ldots, n_i\}$, as long as the mean $\{\bar{y}_0^P, \bar{y}_1^P, \bar{y}_2^P, \bar{y}_0^S, \bar{y}_1^S, \bar{y}_2^S\}$ is fixed, the constraint MLEs are fixed as well. The acceptance region in (4.9) is general ellipsoid. We completes the proof of Theorem 7. ■

Then the intersection hypothesis $H_{e_1e_2}$ is rejected if $\lambda_{e_1e_2} < c_{e_1e_2}$ where the critical value $c_{e_1e_2}$ is determined by the equation following Theorem 7:

$$\inf_{\boldsymbol{\theta} \in H_{e_1e_2}} P_{\boldsymbol{\theta},\rho}(\lambda_{e_1e_2} \geq c_{e_1e_2}) = P_{\boldsymbol{\theta}=\boldsymbol{\delta},\rho}(\lambda_{e_1e_2} \geq c_{e_1e_2}) \geq P_{\boldsymbol{\theta}=\boldsymbol{\delta},\rho=0}(\lambda_{e_1e_2} \geq c_{e_1e_2}) = 1 - \alpha.$$

The least favorable configuration for $\lambda_{e_1e_2}$ is $\boldsymbol{\theta} = \boldsymbol{\delta}$, which is proved as follows.

Proof of Least Favorable Configuration. Without lose of generality, we consider the simple case with $\boldsymbol{\delta} = \mathbf{0}$. According to (4.8), the null hypothesis $H_{e_1e_2}$ is rejected for small

values of $\lambda_{e_1e_2}$ or

$$(\lambda_{e_1e_2})^{2/\sum n_i} = \frac{\sum_{ij}(y_{ij}^P - \bar{y}_i^P)^2 \sum_{ij}(y_{ij}^S - \bar{y}_i^S)^2}{\sum_{ij}(y_{ij}^P - \tilde{y}_i^P)^2 \sum_{ij}(y_{ij}^S - \tilde{y}_i^S)^2}$$
$$= \frac{\sum_{ij}(y_{ij}^P - \bar{y}_i^P)^2}{\sum_{ij}(y_{ij}^P - \bar{y}_i^P)^2 + \sum_i n_i(\bar{y}_i^P - \tilde{y}_i^P)^2} \cdot \frac{\sum_{ij}(y_{ij}^S - \bar{y}_i^S)^2}{\sum_{ij}(y_{ij}^S - \bar{y}_i^S)^2 + \sum_i n_i(\bar{y}_i^S - \tilde{y}_i^S)^2}$$

It is equivalent to reject $H_{e_1e_2}$ for large values of

$$\frac{\sum_i n_i(\bar{y}_i^P - \tilde{y}_i^P)^2}{\sum_{ij}(y_{ij}^P - \bar{y}_i^P)^2} \cdot \frac{\sum_i n_i(\bar{y}_i^S - \tilde{y}_i^S)^2}{\sum_{ij}(y_{ij}^S - \bar{y}_i^S)^2}$$

As $\sum_{ij}(y_{ij}^E - \bar{y}_i^E)^2/\sigma_E^2$ follows chi-square distribution and is independent of $\sum_i n_i(\bar{y}_i^E - \tilde{y}_i^E)^2$ for $E = P$ or S, respectively, the least favorable configuration does not depend on it.

Let $(\boldsymbol{\eta}^P, \boldsymbol{\eta}^S) = (\eta_0^P, \eta_1^P, \eta_2^P, \eta_0^S, \eta_1^S, \eta_2^S) \in H_{e_1e_2}$ be any configuration in the null space and let

$$\begin{pmatrix} W_i^P \\ W_i^S \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_P^2/n_i & \rho\sigma_P\sigma_S/n_i \\ \rho\sigma_P\sigma_S/n_i & \sigma_S^2/n_i \end{pmatrix}\right) \qquad \text{for } i = 0, 1, 2.$$

Thus according to (4.7),

$$\begin{pmatrix} \bar{Y}_i^P \\ \bar{Y}_i^S \end{pmatrix} \overset{\text{d}}{=} \begin{pmatrix} W_i^P \\ W_i^S \end{pmatrix} + \begin{pmatrix} \mu_i^P \\ \mu_i^S \end{pmatrix}$$

For $\boldsymbol{\mu} = \boldsymbol{\eta}$,

$$\sum_i n_i(\bar{y}_i^P - \tilde{y}_i^P)^2 \cdot \sum_i n_i(\bar{y}_i^S - \tilde{y}_i^S)^2$$
$$\overset{\text{d}}{=} \sum_i n_i[W_i^P + \eta_i^P - (\mathbf{W}^P + \boldsymbol{\eta}^P)_i^*]^2 \cdot \sum_i n_i[W_i^S + \eta_i^S - (\mathbf{W}^S + \boldsymbol{\eta}^S)_i^*]^2$$

and for $\boldsymbol{\mu} = \mathbf{0}$,

$$\sum_i n_i(\bar{y}_i^P - \tilde{y}_i^P)^2 \cdot \sum_i n_i(\bar{y}_i^S - \tilde{y}_i^S)^2 \overset{\text{d}}{=} \sum_i n_i(W_i^P - W_i^{P*})^2 \cdot \sum_i n_i(W_i^S - W_i^{S*})^2$$

with $(\mathbf{W}^E + \boldsymbol{\eta}^E)_i^*$ and W_i^{E*} the ith component of the isotonic regression of $\mathbf{W}^E + \boldsymbol{\eta}^E$ and $\mathbf{W}^E$ with weight n_i, for $E = P$ or S, respectively. For more detailed discussion on isotonic regression, we refer to Robertson et al. (1988).

Because $\mathbf{W}^* + \boldsymbol{\eta} \in H_{e_1e_2}$ and the definition of isotonic regression, we have the following inequality,

$$\begin{aligned}
&\sum_i n_i[W_i^P + \eta_i^P - (\mathbf{W}^P + \boldsymbol{\eta}^P)_i^*]^2 \cdot \sum_i n_i[W_i^S + \eta_i^S - (\mathbf{W}^S + \boldsymbol{\eta}^S)_i^*]^2 \\
\leq &\sum_i n_i[W_i^P + \eta_i^P - (W_i^{P*} + \eta_i^P)]^2 \cdot \sum_i n_i[W_i^S + \eta_i^S - (W_i^{S*} + \eta_i^S)]^2 \\
= &\sum_i n_i(W_i^P - W_i^{P*})^2 \cdot \sum_i n_i(W_i^S - W_i^{S*})^2
\end{aligned}$$

Hence $\boldsymbol{\mu} = \mathbf{0} \in H_{e_1e_2}$ is least favorable for $\sum_i n_i(\bar{y}_i^P - \tilde{y}_i^P)^2 \cdot \sum_i n_i(\bar{y}_i^S - \tilde{y}_i^S)^2$. The least favorable configuration for a general $\boldsymbol{\delta}$ is simply $\boldsymbol{\theta} = \boldsymbol{\delta}$. ■

When $\rho = 0$, the distribution of $\lambda_{e_1e_2}$ can be generated by Monte-Carlo simulation of random variables

$$\mathbf{Y}_{ij}^{(t)} \overset{\text{i.i.d.}}{\sim} N\left(\begin{pmatrix} \delta^P \\ \\ \delta^S \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \\ 0 & 1 \end{pmatrix}\right), \quad t = 1, 2, \ldots, r.$$

Since $\lambda_{e_1e_2}$ is the ratio of variance estimates, the unknown values of σ_P and σ_S do not contribute to the distribution. If a total number of r Monte-Carlo samples are generated, then the critical value $c_{e_1e_2}$ can be simply estimated by the lower αth quantile of the sample statistics $\{\lambda_{e_1e_2}^{(t)} : t = 1, \ldots, r\}$.

4.3.2 Consonance Adjustment

As likelihood ratio test does not rely on the union-intersection method, the procedure may not satisfy the logical consistency requirement of consonance. The formal definition of consonance and a general solution to the consistency problem in multiple hypotheses testing has been discussed in Chapter 2. If an intersection hypothesis was rejected by a consonant test, at least one of the component hypotheses should have been rejected as well. Unfortunately, likelihood ratio test does not have such property.

To overcome this kind of inconsistency, we provided in Chapter 3 a consonant likelihood ratio type closed test for the problem of multiple comparisons with a control in dose-response study. When multiple doses are combined with multiple endpoints, the test procedure becomes more complicated. In this chapter, we propose a consonance adjustment procedure for partition testing with multiple endpoints. With this consonance adjustment steps, the likelihood ratio test becomes a powerful tool to evaluate the efficacy of treatment effects.

Notice that in Table 4.1, null hypothesis $H_{e_1e_2}$ can be expressed as intersections $H_1^{e_1} \cap H_2^{e_2}$ with

$$H_i^{e_i} = \begin{cases} \{\theta_i^P \leq \delta^P\} & \text{if } e_i = 1 \\ \{\theta_i^S \leq \delta^S\} & \text{if } e_i = 2\,, \\ \{-\infty < \theta_i^P < \infty \text{ and } -\infty < \theta_i^S < \infty\} & \text{if } e_i = 3 \end{cases} \qquad i = 1, 2. \tag{4.11}$$

If none of e_1 or e_2 equals 3, we reject $H_{e_1e_2}$ when the likelihood ratio statistic $\lambda_{e_1e_2} \leq c'_{e_1e_2}$ and at least one of the component hypotheses $H_1^{e_1}$ and $H_2^{e_2}$ is rejected. Otherwise $H_{e_1e_2}$ is not an intersection hypothesis and no adjustment is necessary. The critical constant $c'_{e_1e_2}$ can be calculated by Monte-Carlo method as well. The final inference is reached by collating the results following the steps introduced in Section 4.2.1.

4.4 Likelihood Ratio Test and Consonance Adjustment: Advanced Case

4.4.1 Primary-Secondary Endpoints with 3 Doses and a Control

Consider a single primary endpoint P and a single secondary endpoint S with $k = 3$ doses. The response of the jth patient in dose group i is denoted by $\mathbf{Y}_{ij}$, which follows the same model as (4.7). Use $i = 0$ to denote the control group and ρ, σ_P, σ_S are unknown constants.

The only assumption for ρ is that $\rho \in (-1/(m-1), 1)$ such that the variance covariance matrix is positive definite, where $m = 2$ is the number of endpoints in this example.

Following the path partition of Section 4.2.1, for each dose i, the parameter space is partitioned into $\Theta_i^1 = H_i^P$, $\Theta_i^2 = H_i^{\star S} = H_i^S \backslash H_i^P$ and $\Theta_i^3 = \{\theta_i^P > 0 \text{ and } \theta_i^S > 0\}$ as (4.3). The disjointness partition results in $3^3 - 1 = 26$ null hypotheses denoted as $\Theta_{e_1e_2e_3} = \cap_{i=1}^3 \Theta_i^{e_i}$ and listed in Table 4.2. Again, only the part of null hypothesis states $\theta_i^E \leq \delta^E$ are tested. The null hypotheses actually tested are listed in Table 4.2 as well. The likelihood ratio statistics are defined similarly to (4.8). Each null hypothesis $H_{e_1e_2e_3}$ is an intersection $H_1^{e_1} \cap H_2^{e_2} \cap H_3^{e_3}$, where $H_i^{e_i}$ is defined in (4.11). In the consonance adjustment, we view $H_{e_1e_2e_3}$ as an intersection of three component hypotheses $H_1^{e_1} \cap H_2^{e_2}$, $H_1^{e_1} \cap H_3^{e_3}$ and $H_2^{e_2} \cap H_3^{e_3}$. Details on calculating the critical constants in the procedure is illustrated in Section 4.4.2.

Finally, the efficacy of dose i in primary endpoint is established if H_i^P is rejected, which means all the natural partitions $\Theta_{e_1e_2e_3}$ ($H_{e_1e_2e_3}$) with $e_i = 1$ are rejected. For example, H_2^P will be rejected if all the natural partitions $\Theta_{e_11e_3}$ for any e_1, e_3 values are rejected. In order to show the efficacy of dose i in secondary endpoint, both H_i^P and $H_i^{\star S}$ should be rejected, which means all the natural partitions $\Theta_{e_1e_2e_3}$ with $e_i \leq 2$ need to be rejected. In the example of $i = 2$ (second dose), all the natural partitions $\Theta_{e_11e_3}$ and $\Theta_{e_12e_3}$ for any e_1, e_3 values should be rejected to show efficacy of the second dose in secondary endpoint. By collating the result, we have a consistent multiple testing procedure to evaluate the efficacy of the three doses in the primary and secondary endpoints.

4.4.2 Notes for critical constants calculation

We illustrate the calculation of critical constants in the example of 3 doses with primary and secondary endpoints. Suppose $H_{e_1e_2e_3}$ is an intersection hypothesis as defined in

Hypotheses of interest	θ_1^P	θ_1^S	θ_2^P	θ_2^S	θ_3^P	θ_3^S	Hypotheses tested	θ_1^P	θ_1^S	θ_2^P	θ_2^S	θ_3^P	θ_3^S
Θ_{111}	$\leq \delta^P$	-	$\leq \delta^P$	-	$\leq \delta^P$	-	H_{111}	$\leq \delta^P$	-	$\leq \delta^P$	-	$\leq \delta^P$	-
Θ_{112}	$\leq \delta^P$	-	$\leq \delta^P$	-	$> \delta^P$	$\leq \delta^S$	H_{112}	$\leq \delta^P$	-	$\leq \delta^P$	-	-	$\leq \delta^S$
Θ_{121}	$\leq \delta^P$	-	$> \delta^P$	$\leq \delta^S$	$\leq \delta^P$	-	H_{121}	$\leq \delta^P$	-	-	$\leq \delta^S$	$\leq \delta^P$	-
Θ_{211}	$> \delta^P$	$\leq \delta^S$	$\leq \delta^P$	-	$\leq \delta^P$	-	H_{211}	-	$\leq \delta^S$	$\leq \delta^P$	-	$\leq \delta^P$	-
Θ_{113}	$\leq \delta^P$	-	$\leq \delta^P$	-	$> \delta^P$	$> \delta^S$	H_{113}	$\leq \delta^P$	-	$\leq \delta^P$	-	-	-
Θ_{131}	$\leq \delta^P$	-	$> \delta^P$	$> \delta^S$	$\leq \delta^P$	-	H_{131}	$\leq \delta^P$	-	-	-	$\leq \delta^P$	-
Θ_{311}	$> \delta^P$	$> \delta^S$	$\leq \delta^P$	-	$\leq \delta^P$	-	H_{311}	-	-	$\leq \delta^P$	-	$\leq \delta^P$	-
Θ_{122}	$\leq \delta^P$	-	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$\leq \delta^S$	H_{122}	$\leq \delta^P$	-	-	$\leq \delta^S$	-	$\leq \delta^S$
Θ_{212}	$> \delta^P$	$\leq \delta^S$	$\leq \delta^P$	-	$> \delta^P$	$\leq \delta^S$	H_{212}	-	$\leq \delta^S$	$\leq \delta^P$	-	-	$\leq \delta^S$
Θ_{221}	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$\leq \delta^S$	$\leq \delta^P$	-	H_{221}	-	$\leq \delta^S$	-	$\leq \delta^S$	$\leq \delta^P$	-
Θ_{123}	$\leq \delta^P$	-	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$> \delta^S$	H_{123}	$\leq \delta^P$	-	-	$\leq \delta^S$	-	-
Θ_{132}	$\leq \delta^P$	-	$> \delta^P$	$> \delta^S$	$> \delta^P$	$\leq \delta^S$	H_{132}	$\leq \delta^P$	-	-	-	-	$\leq \delta^S$
Θ_{213}	$> \delta^P$	$\leq \delta^S$	$\leq \delta^P$	-	$> \delta^P$	$> \delta^S$	H_{213}	-	$\leq \delta^S$	$\leq \delta^P$	-	-	-
Θ_{312}	$> \delta^P$	$> \delta^S$	$\leq \delta^P$	-	$> \delta^P$	$\leq \delta^S$	H_{312}	-	-	$\leq \delta^P$	-	-	$\leq \delta^S$
Θ_{231}	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$> \delta^S$	$\leq \delta^P$	-	H_{231}	-	$\leq \delta^S$	-	-	$\leq \delta^P$	-
Θ_{321}	$> \delta^P$	$> \delta^S$	$> \delta^P$	$\leq \delta^S$	$\leq \delta^P$	-	H_{321}	-	-	-	$\leq \delta^S$	$\leq \delta^P$	-
Θ_{133}	$\leq \delta^P$	-	$> \delta^P$	$> \delta^S$	$> \delta^P$	$> \delta^S$	H_{133}	$\leq \delta^P$	-	-	-	-	-
Θ_{313}	$> \delta^P$	$> \delta^S$	$\leq \delta^P$	-	$> \delta^P$	$> \delta^S$	H_{313}	-	-	$\leq \delta^P$	-	-	-
Θ_{331}	$> \delta^P$	$> \delta^S$	$> \delta^P$	$> \delta^S$	$\leq \delta^P$	-	H_{331}	-	-	-	-	$\leq \delta^P$	-
Θ_{222}	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$\leq \delta^S$	H_{222}	-	$\leq \delta^S$	-	$\leq \delta^S$	-	$\leq \delta^S$
Θ_{223}	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$> \delta^S$	H_{223}	-	$\leq \delta^S$	-	$\leq \delta^S$	-	-
Θ_{232}	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$> \delta^S$	$> \delta^P$	$\leq \delta^S$	H_{232}	-	$\leq \delta^S$	-	-	-	$\leq \delta^S$
Θ_{322}	$> \delta^P$	$> \delta^S$	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$\leq \delta^S$	H_{322}	-	-	-	$\leq \delta^S$	-	$\leq \delta^S$
Θ_{233}	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$> \delta^S$	$> \delta^P$	$> \delta^S$	H_{233}	-	$\leq \delta^S$	-	-	-	-
Θ_{323}	$> \delta^P$	$> \delta^S$	$> \delta^P$	$\leq \delta^S$	$> \delta^P$	$> \delta^S$	H_{323}	-	-	-	$\leq \delta^S$	-	-
Θ_{332}	$> \delta^P$	$> \delta^S$	$> \delta^P$	$> \delta^S$	$> \delta^P$	$\leq \delta^S$	H_{332}	-	-	-	-	-	$\leq \delta^S$

Table 4.2: The natural partition of primary-secondary endpoints with 3 doses and a control.

Table 4.2 with $e_1, e_2, e_3 \leq 2$. It is obvious that $H_{e_1e_2e_3} = H_1^{e_1} \cap H_2^{e_2} \cap H_3^{e_3}$ is a "three-way" intersection hypothesis. Similarly, we can define "two-way" intersection hypotheses such as $H_1^{e_1} \cap H_2^{e_2}$, $H_1^{e_1} \cap H_3^{e_3}$ and $H_2^{e_2} \cap H_3^{e_3}$. The "three-way" intersection hypothesis $H_{e_1e_2e_3}$ is the intersection of any two "two-way" intersection hypotheses. We also notice that the "two-way" intersection hypotheses can be labeled as $H_{e_1e_23}$, $H_{e_13e_3}$ and $H_{3e_2e_3}$ which appears in Table 4.2. The elementary components of these hypotheses, namely, $H_1^{e_1}$, $H_2^{e_2}$ and $H_3^{e_3}$ are called elementary hypotheses. They can be labeled as H_{e_133}, H_{3e_23} and H_{33e_3}, respectively.

The consonance adjustment works in a step-wise fashion. A "three-way" intersec-

tion hypothesis is rejected providing that at least two "two-way" intersection hypotheses are rejected (at most one "two-way" intersection hypotheses is retained). A "two-way" intersection hypothesis can be rejected only if at least one of the elementary hypotheses is rejected.

The computation of critical constants follows reversed steps. For simplicity, we denote the likelihood ratio statistics and critical constants of a hypothesis as λ and c, respectively, with the same index as the hypothesis. Suppose r Monte-Carlo samples are generated following the null distribution. The likelihood ratio statistics for each intersection hypothesis in Table 4.2 are calculated for all r samples. First consider all the elementary hypotheses H_{133}, H_{233}, H_{313}, H_{323}, H_{331} and H_{332}. The critical constants for these hypotheses are just the lower α quantile of the likelihood ratio statistics λ (among r Monte-Carlo samples). In the second step we consider "two-way" intersection hypotheses, for example, H_{132}. We adjust its likelihood ratio statistic $\lambda_{132}^{(t)}$ in the t-th Monte-Carlo sample such that if both elementary hypotheses H_{133} and H_{332} are retained by $\lambda_{133}^{(t)} \geq c_{133}$ and $\lambda_{332}^{(t)} \geq c_{332}$, we set $\lambda_{132}^{(t)} = \infty$ since we should not reject H_{132} in this case. After the adjustment is completed for $t = 1, \ldots, r$ samples, the critical constant c_{132} is calculated as the lower α quantile of the truncated sample likelihood ratio statistics. This step is executed for all the "two-way" intersection hypotheses. Proceed to the next step, we make adjustment for "three-way" intersection hypotheses. Take H_{121} for example, it is the intersection of three "two-way" intersections $H_{123} \cap H_{131} \cap H_{321}$. Consider the likelihood ratio statistic $\lambda_{121}^{(t)}$ in the t-th Monte-Carlo sample. It will be set to ∞ if two or more of the "two-way" intersections are retained, namely, $\lambda_{123}^{(t)} \geq c_{123}$, $\lambda_{131}^{(t)} \geq c_{131}$ or $\lambda_{321}^{(t)} \geq c_{321}$. Notice that if $\lambda_{123}^{(t)} < c_{123}$, it indicates that $\lambda_{123}^{(t)}$ has not been set to ∞ in the previous step, thus at least one of the elementary hypotheses H_{133} and H_{323} has already been rejected. Same conclusion applies to $\lambda_{131}^{(t)}$ and $\lambda_{321}^{(t)}$. When the adjustment is completed for all the r samples, the critical constant c_{121} is calculated as the lower α quantile of the sample likelihood ratio statistics $\lambda_{121}^{(t)}$ for $t = 1, \ldots, r$.

4.4.3 Simulation study and clinical trial example

The performance of the consonant likelihood ratio test (CLRT) is compared with two existing procedures, namely, the decision path partition testing of Liu and Hsu (2009) (DP) with critical value calculation based on joint distribution within endpoint and Bonferroni inequality across endpoints, and the tree-gatekeeping procedure of Dmitrienko et al. (2007b) (GK). The simulation study is carried out comparing three dose groups to a placebo group with 50 patients in each group. The correlation ρ between the primary and secondary endpoint is assumed unknown. Four different scenarios of true mean matrix are considered as presented in Table 4.3 to represent different dose-response relationships with respect to primary and secondary endpoints. A higher value indicates stronger effect. Two values of ρ are studied to evaluate the effect of true correlation among endpoints.

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
Dose	1	2	3	1	2	3	1	2	3	1	2	3
P	0.65	0.65	0.65	0.5	0.65	0.8	0.5	0.65	0.8	0	0.65	0.65
S	0.65	0.65	0.65	0.5	0.65	0.8	0.8	0.65	0.5	0.65	0	0.65

Table 4.3: Scenarios of mean matrices for power comparison

The simulation result is shown in Table 4.4 based on $1,000,000$ runs with FWER controlled strongly at 0.025. We see that the CLRT procedure has a significant power advantage in the "central direction" when all the null hypotheses are false (all dose levels are efficacious in both endpoints). However, if most of the null hypotheses are in fact true, then the t test based DP procedures are preferred.

We also notice that the performance comparison among the procedure is not affected by the correlation ρ among endpoints. CLRT procedure has the highest power in Scenarios 1, 2

and 3 for all the correlation values experimented. Besides, the type I errors in Scenario 4 drops as ρ changes to negative values.

		correlation $\rho = 0.3$											
		Scenario 1			Scenario 2			Scenario 3			Scenario 4		
Dose $\rightarrow$		1	2	3	1	2	3	1	2	3	1	2	3
CLRT	P	0.869	0.869	0.868	0.677	0.856	0.954	0.670	0.855	0.955	0.014	0.814	0.755
	S	0.772	0.771	0.771	0.499	0.752	0.915	0.653	0.752	0.653	0.014	0.014	0.606
DP	P	0.850	0.850	0.850	0.662	0.846	0.951	0.632	0.836	0.953	0.015	0.823	0.806
	S	0.745	0.745	0.745	0.489	0.734	0.906	0.617	0.722	0.609	0.014	0.012	0.661
GK	P	0.792	0.792	0.790	0.534	0.792	0.939	0.533	0.792	0.939	0.009	0.793	0.793
	S	0.695	0.695	0.692	0.400	0.690	0.895	0.518	0.684	0.611	0.009	0.016	0.654
		correlation $\rho = -0.3$											
		Scenario 1			Scenario 2			Scenario 3			Scenario 4		
Dose $\rightarrow$		1	2	3	1	2	3	1	2	3	1	2	3
CLRT	P	0.869	0.869	0.869	0.678	0.857	0.955	0.671	0.857	0.955	0.012	0.814	0.758
	S	0.746	0.747	0.746	0.425	0.724	0.910	0.634	0.724	0.634	0.007	0.007	0.553
DP	P	0.850	0.850	0.850	0.662	0.845	0.951	0.632	0.839	0.952	0.014	0.823	0.805
	S	0.713	0.713	0.713	0.413	0.700	0.901	0.595	0.689	0.588	0.008	0.007	0.614
GK	P	0.794	0.793	0.794	0.535	0.790	0.938	0.535	0.793	0.938	0.009	0.793	0.788
	S	0.656	0.655	0.656	0.316	0.649	0.886	0.497	0.643	0.584	0.005	0.009	0.601

Table 4.4: Power comparison for four scenarios with true difference specified in Table 4.3.

The performance of the three procedures is also displayed in Figure 4.3 for the case $\rho = 0.3$. Since the relative performance remains unchanged for different ρ values, we will not show the graphs for other ρ values.

A clinical trial example was studied in Dmitrienko et al. (2003) and Liu and Hsu (2009) which evaluate the effect of a hypertension drug. Three doses are compared to a placebo in two endpoints. The primary endpoint is the reduction in systolic blood pressure (SBP) and the secondary endpoint is the reduction in diastolic blood pressure (DBP). The mean and standard

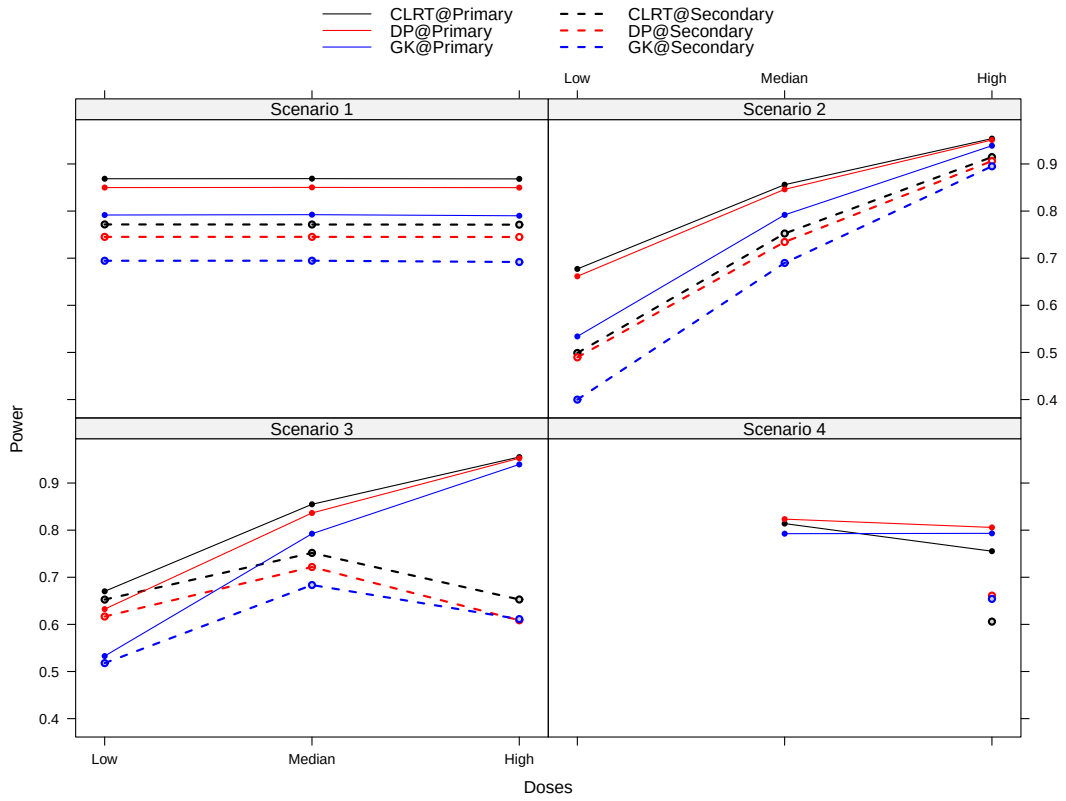

Figure 4.3: The performance comparison when the correlation between primary endpoint and secondary endpoint is 0.3.

deviation for each dose and endpoint combination are listed in Table 4.5.

Since both endpoints measure the reductions, a lower level indicates better treatment. If the FWER is controlled strongly at the 0.05 level, the gatekeeping procedure identifies the median dose and high dose as effective in the primary endpoint and only the median dose is efficacious in the secondary endpoint. Both the decision path partition testing procedure and the consonant likelihood ratio test procedure reject all null hypotheses and establish efficacy for all dose-endpoints combinations.

Treatment group	n	Change in SBP (mmHg)		Change in DBP (mmHg)	
		Mean	Standard Deviation	Mean	Standard Deviation
Placebo	29	4.02	7.85	2.31	5.88
Low dose	23	-2.16	8.71	-0.67	5.51
Median dose	24	-5.03	11.01	-3.18	6.89
High dose	24	-2.60	8.96	-1.44	6.15

Table 4.5: Results of a dose-finding study in patients with hypertension.

4.4.4 Co-Primary Endpoints with 2 Doses and a Control

The consonance adjustment procedure discussed above can be easily extended to cases besides the hierarchical ordered endpoints. Incorporate with the extended partitioning principles introduced in Section 4.2.2 and 4.2.3, we can apply the consonant likelihood ratio test in alternative primary and co-primary endpoint cases. Taking the co-primary endpoint for example, suppose there are two co-primary endpoints P1, P2 and a single secondary endpoint S for $k = 2$ doses. The response of the jth patient in dose group i follows the model

$$\mathbf{Y}_{ij} \overset{\text{i.i.d.}}{\sim} N\left(\begin{pmatrix} \mu_i^{P1} \\ \mu_i^{P2} \\ \mu_i^{S} \end{pmatrix}, \begin{pmatrix} \sigma_{P1}^2 & \rho\sigma_{P1}\sigma_{P2} & \rho\sigma_{P1}\sigma_S \\ \rho\sigma_{P1}\sigma_{P2} & \sigma_{P2}^2 & \rho\sigma_{P2}\sigma_S \\ \rho\sigma_{P1}\sigma_S & \rho\sigma_{P2}\sigma_S & \sigma_S^2 \end{pmatrix}\right), \quad i = 0, 1, 2,\ j = 1, 2, \cdots, n_i;$$

where ρ, σ_{P1}, σ_{P2} and σ_S are unknown constants. The condition $\rho \in (-1/2, 1)$ is imposed here to make the variance-covariance matrix positive definite.

Following the path partition of Section 4.2.3, for each dose i, the parameter space is partitioned into $\Theta_i = \{\Theta_i^{P1}, \Theta_i^{P2}, \Theta_i^{P1P2}, \Theta_i^{\star S}, \Theta_i^{A}\}$ as defined in (4.6). They can be ordered and relabeled as

$$\Theta_i^1 = \Theta_i^{P1P2},\ \Theta_i^2 = \Theta_i^{P1},\ \Theta_i^3 = \Theta_i^{P2},\ \Theta_i^4 = \Theta_i^{\star S},\ \Theta_i^5 = \Theta_i^{A}.$$

The disjointness partition results in $5^2 - 1 = 24$ disjoint null hypotheses. We conveniently label them as $\Theta_{e_1e_2} = \cap_{i=1}^{2} \Theta_i^{e_i}$. The complete list of the 24 null hypotheses of interest can be found in Table 4.6. Also included in Table 4.6 are the hypotheses actually tested eliminating the parts stating $\theta_i^E > \delta^E$.

Hypotheses of interest	θ_1^{P1}	θ_1^{P2}	θ_1^S	θ_2^{P1}	θ_2^{P2}	θ_2^S	Hypotheses tested	θ_1^{P1}	θ_1^{P2}	θ_1^S	θ_2^{P1}	θ_2^{P2}	θ_2^S
Θ_{11}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	H_{11}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-
Θ_{12}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	$\leq \delta^{P1}$	$> \delta^{P2}$	-	H_{12}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	$\leq \delta^{P1}$	-	-
Θ_{13}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	$> \delta^{P1}$	$\leq \delta^{P2}$	-	H_{13}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	-	$\leq \delta^{P2}$	-
Θ_{21}	$\leq \delta^{P1}$	$> \delta^{P2}$	-	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	H_{21}	$\leq \delta^{P1}$	-	-	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-
Θ_{31}	$> \delta^{P1}$	$\leq \delta^{P2}$	-	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	H_{31}	-	$\leq \delta^{P2}$	-	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-
Θ_{14}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	H_{14}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	-	-	$\leq \delta^S$
Θ_{41}	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	H_{41}	-	-	$\leq \delta^S$	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-
Θ_{15}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	$> \delta^{P1}$	$> \delta^{P2}$	$> \delta^S$	H_{15}	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	-	-	-
Θ_{51}	$> \delta^{P1}$	$> \delta^{P2}$	$> \delta^S$	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-	H_{51}	-	-	-	$\leq \delta^{P1}$	$\leq \delta^{P2}$	-
Θ_{22}	$\leq \delta^{P1}$	$> \delta^{P2}$	-	$\leq \delta^{P1}$	$> \delta^{P2}$	-	H_{22}	$\leq \delta^{P1}$	-	-	$\leq \delta^{P1}$	-	-
Θ_{23}	$\leq \delta^{P1}$	$> \delta^{P2}$	-	$> \delta^{P1}$	$\leq \delta^{P2}$	-	H_{23}	$\leq \delta^{P1}$	-	-	-	$\leq \delta^{P2}$	-
Θ_{32}	$> \delta^{P1}$	$\leq \delta^{P2}$	-	$\leq \delta^{P1}$	$> \delta^{P2}$	-	H_{32}	-	$\leq \delta^{P2}$	-	$\leq \delta^{P1}$	-	-
Θ_{33}	$> \delta^{P1}$	$\leq \delta^{P2}$	-	$> \delta^{P1}$	$\leq \delta^{P2}$	-	H_{33}	-	$\leq \delta^{P2}$	-	-	$\leq \delta^{P2}$	-
Θ_{24}	$\leq \delta^{P1}$	$> \delta^{P2}$	-	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	H_{24}	$\leq \delta^{P1}$	-	-	-	-	$\leq \delta^S$
Θ_{42}	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	$\leq \delta^{P1}$	$> \delta^{P2}$	-	H_{42}	-	-	$\leq \delta^S$	$\leq \delta^{P1}$	-	-
Θ_{34}	$> \delta^{P1}$	$\leq \delta^{P2}$	-	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	H_{34}	-	$\leq \delta^{P2}$	-	-	-	$\leq \delta^S$
Θ_{43}	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	$> \delta^{P1}$	$\leq \delta^{P2}$	-	H_{43}	-	-	$\leq \delta^S$	-	$\leq \delta^{P2}$	-
Θ_{44}	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	H_{44}	-	-	$\leq \delta^S$	-	-	$\leq \delta^S$
Θ_{25}	$\leq \delta^{P1}$	$> \delta^{P2}$	-	$> \delta^{P1}$	$> \delta^{P2}$	$> \delta^S$	H_{25}	$\leq \delta^{P1}$	-	-	-	-	-
Θ_{52}	$> \delta^{P1}$	$> \delta^{P2}$	$> \delta^S$	$\leq \delta^{P1}$	$> \delta^{P2}$	-	H_{52}	-	-	-	$\leq \delta^{P1}$	-	-
Θ_{35}	$> \delta^{P1}$	$\leq \delta^{P2}$	-	$> \delta^{P1}$	$> \delta^{P2}$	$> \delta^S$	H_{35}	-	$\leq \delta^{P2}$	-	-	-	-
Θ_{53}	$> \delta^{P1}$	$> \delta^{P2}$	$> \delta^S$	$> \delta^{P1}$	$\leq \delta^{P2}$	-	H_{53}	-	-	-	-	$\leq \delta^{P2}$	-
Θ_{45}	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	$> \delta^{P1}$	$> \delta^{P2}$	$> \delta^S$	H_{45}	-	-	$\leq \delta^S$	-	-	-
Θ_{54}	$> \delta^{P1}$	$> \delta^{P2}$	$> \delta^S$	$> \delta^{P1}$	$> \delta^{P2}$	$\leq \delta^S$	H_{54}	-	-	-	-	-	$\leq \delta^S$

Table 4.6: The natural partition of co-primary endpoints with 2 doses and a control.

In collating the test result of each hypothesis $H_{e_1e_2}$, we claim dose i is efficacious in primary endpoint P1 if the elementary hypothesis H_i^{P1} is rejected, which includes the rejections of all natural partitions $H_{e_1e_2}$ with $e_i = 1$ or 2. Dose i is claimed to be efficacious in primary endpoint P2 if H_i^{P2} is rejected, which includes the rejections of all natural partitions $H_{e_1e_2}$ with

$e_i = 1$ or 3. To claim dose i is efficacious in the secondary endpoint S needs the rejection of both H_i^{P1} and H_i^{P2} as well as $H_i^{\star S}$, which includes the rejections of all natural partitions $H_{e_1e_2}$ with $e_i \leq 4$.

The consonance likelihood ratio test serves as intermediate steps to test each intersection hypothesis $H_{e_1e_2}$ using likelihood ratio statistics and adjusted for consistency. Specifically, if partition hypothesis $H_{e_1e_2}$ (defined in Table 4.6) has indices $e_1, e_2 \in \{2, 3, 4\}$, then $H_{e_1e_2} = H_1^{e_1} \cap H_2^{e_2}$ and the rejection of $H_{e_1e_2}$ requires the rejection of at least one of the elementary hypothesis $H_1^{e_1}$ or $H_2^{e_2}$, which is in fact H_{e_15} and H_{5e_2}. For example, H_{24} should be rejected only if at least one of H_{25} or H_{54} is rejected.

If e_1, e_2 are 1 and 5, for example, H_{15}. It is rejected if at least one of H_{25} or H_{35} is rejected.

If one of e_1, e_2 is 1 and the other is not 5. For example, H_{14} should be rejected if there is at most one acceptance among H_{25}, H_{35} and H_{54} since $H_{1e_2} = H_{25} \cap H_{35} \cap H_{5e_2}$.

If both $e_1 = e_2 = 1$, H_{11} is rejected if at most one acceptance among H_{25}, H_{35}, H_{52} and H_{53}.

If one of e_1, e_2 is 5 and the other is not 1, the partition is an elementary hypothesis and no consonance adjustment is needed. The hypothesis H_{55} does not need to be tested.

The critical values of the consonant likelihood ratio test is calculated in a reversed order. Suppose r Monte-Carlo samples are generated following the null distribution that there is no treatment effect for any dose on any endpoint. The likelihood ratio statistics for each intersection hypothesis in Table 4.6 are calculated. First consider all the elementary hypotheses H_{25}, H_{35}, H_{45}, H_{52}, H_{53} and H_{54}. The critical constants for these hypotheses are just the lower α quantile of the likelihood ratio statistics among r Monte-Carlo samples.

In the second step we consider hypotheses in the right panel of Table 4.6 which has two components. They are H_{15}, H_{51} and all $H_{e_1e_2}$ with $e_1, e_2 \in \{2, 3, 4\}$. The likelihood

ratio statistics of those hypotheses are set to ∞ if both component elementary hypotheses are retained by previous step. Critical constants are lower α quantiles of the adjusted Monte-Carlo likelihood ratio statistics.

Next, the intersection of three elementary hypotheses are considered. Namely, H_{1e_2} and H_{e_11} for $e_1, e_2 \in \{2, 3, 4\}$. The critical constant for H_{1e_2} is the lower α quantiles of the adjusted likelihood ratio statistics, which has been set to ∞ if more than one acceptance among the components H_{15}, H_{2e_2} and H_{3e_2} (or H_{51}, H_{e_12} and H_{e_13} for H_{e_11}).

Finally, H_{11} is considered. The critical constant is found after Monte-Carlo sample likelihood ratio statistics has be set to ∞ for those samples with more than one acceptance among H_{12}, H_{13}, H_{21} and H_{31}, which are intersection hypotheses with three components. Notice that H_{11} is the intersection of any two hypotheses among H_{12}, H_{13}, H_{21} and H_{31}.

Remark. The covariance matrix is assumed to have a compound symmetry structure in the co-primary endpoint problems. However, this is not a necessary condition such that Theorem 1 holds. In fact, as long as the covariance matrix is positive definite, likelihood ratio based test is conservative with independence assumption.

4.5 Discussion and Future Research

In this chapter we propose a new partition testing procedure for multiple endpoint problems, which provides a significant power advantage over the gatekeeping procedure. This advantage is partly due to the generalized decision path partitioning and mostly comes from the asymptotic optimality of likelihood ratio test applied to each disjoint partition hypothesis. When likelihood ratio test is used in multiple hypotheses testing, the logical consistency requirement consonance will be lost. A consonance adjustment for multiple endpoints is proposed in this paper to retain the consistency and reserve the power of likelihood ratio test.

In the framework of partition testing, there are still other options in testing the intersection hypotheses. For example, Liu and Hsu (2009) hints that strong partition (Finner and Strassburger, 2002) can be used to obtain more powerful tests. However, we discovered in our research that the strong partitioning principle (SPP) will not increase the power unless a restricted null space or similar assumptions are used. Even though, the power improvement is marginal.

To illustrate the idea, we will consider the example used in Section 4.3 with two doses and two endpoints. First, the parameter space is partitioned into 9 pieces as displayed in Table 4.1. The strong partitioning principle requires the local level α test $\psi_{e_1e_2}$ for $\Theta_{e_1e_2}$ satisfies the condition $\psi_{e_1e_2} = \min_{\Theta_{ab} \subseteq H_{e_1e_2}} \psi_{ab}$. The partition testing with SPP applied to Θ is supposed to have a different set of critical values than Table 2 in Liu and Hsu (2009), which is showed in Table 4.7. The critical values can be solved recursively.

Partition	SPP test	Rejection Region
Θ_{11}	ψ_{11}	$\{T_1^P > c_1^S$ or $T_2^P > c_1^S\}$
Θ_{12}	ψ_{12}	$\{T_1^P > c_1^S$ or $T_2^S > c_2^S\}$
Θ_{21}	ψ_{21}	$\{T_2^P > c_1^S$ or $T_1^S > c_2^S\}$
Θ_{13}	$\psi_{13} = \min(\psi_{11}, \psi_{12}, \psi_{13})$	$\{T_1^P > c_3^S\}$ and $\{T_1^P > c_1^S$ or $T_2^S > c_2^S\}$ and $\{T_1^P > c_1^S$ or $T_2^P > c_1^S\}$
Θ_{31}	$\psi_{31} = \min(\psi_{11}, \psi_{21}, \psi_{31})$	$\{T_2^P > c_3^S\}$ and $\{T_2^P > c_1^S$ or $T_1^S > c_2^S\}$ and $\{T_1^P > c_1^S$ or $T_2^P > c_1^S\}$
Θ_{22}	ψ_{22}	$\{T_1^S > c_1^S$ or $T_2^S > c_1^S\}$
Θ_{23}	$\psi_{23} = \min(\psi_{21}, \psi_{22}, \psi_{23})$	$\{T_1^S > c_3^S\}$ and $\{T_1^S > c_1^S$ or $T_2^S > c_1^S\}$ and $\{T_2^P > c_1^S$ or $T_1^S > c_2^S\}$
Θ_{32}	$\psi_{32} = \min(\psi_{12}, \psi_{22}, \psi_{32})$	$\{T_2^S > c_3^S\}$ and $\{T_1^S > c_1^S$ or $T_2^S > c_1^S\}$ and $\{T_1^P > c_1^S$ or $T_2^S > c_2^S\}$

Table 4.7: The strong partitioning principle applied to decision path partitioning.

However, after obtaining the least favorable configuration, we can see that the strong partitioning principle results in the same critical constant as the weak partitioning principle used

in Liu and Hsu (2009). Take $\Theta_{13} = \{\theta_1^P \leq \delta^P \text{ and } \theta_2^P > \delta^P \text{ and } \theta_2^S > \delta^S\}$ for example, SPP controls the type-I-error

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta_{13}} P_{\boldsymbol{\theta}}\big(\{T_1^P > c_3\} \text{ and } \{T_1^P > c_1 \text{ or } T_2^S > c_2\} \text{ and } \{T_1^P > c_1 \text{ or } T_2^P > c_1\}\big) \\
=&P_{\theta_1^P=\delta^P,\theta_2^P=\theta_2^S=\infty}\big(\{T_1^P > c_3\} \text{ and } \{T_1^P > c_1 \text{ or } T_2^S > c_2\} \text{ and } \{T_1^P > c_1 \text{ or } T_2^P > c_1\}\big) \\
=&P_{\theta_1^P=\delta^P,\theta_2^P=\theta_2^S=\infty}\big(\{T_1^P > c_3\}\big) = \sup_{\boldsymbol{\theta}\in\Theta_{13}} P_{\boldsymbol{\theta}}\big(\{T_1^P > c_3\}\big)
\end{aligned}$$

which is the same as WPP method.

One possible modification is the restricted null space is assumption (Chuang-Stein et al., 2007). The idea of null space restriction considers $\Theta'_{13} = \{\theta_1^P \leq \delta^P \text{ and } M^P > \theta_2^P > \delta^P \text{ and } M^S > \theta_2^S > \delta^S\}$ as the null parameter space for ψ_{13} ($\psi_{13}^\star$) so that the response of treatment is restricted at a predefined level. It is reasonable since no treatment is expected to have arbitrarily substantial difference to the control. In fact, it is equivalent to change the original hypotheses of interest in (4.2) to

$$H_i^P : \theta_i^P \leq \delta^P \quad \text{vs} \quad A_i^P : \delta^P < \theta_i^P < M^P$$

$$H_i^S : \theta_i^S \leq \delta^S \quad \text{vs} \quad A_i^S : \delta^S < \theta_i^S < M^S$$

With the restricted null space assumption, the least favorable configuration for SPP method on Θ_{13} will occur at $\theta_1^P = \delta^P$, $\theta_2^P = M^P$ and $\theta_2^S = M^S$. The critical constant for SPP will be different from that of WPP.

Chapter 5

Conclusions

Multiple hypotheses testing has been a very active research topic in statistical inferences. The early development on multiple hypotheses testing mainly focus on the closure principle and step-wise closed test procedures with union-intersection setting. During my literature review of multiple hypotheses testing, I noticed that likelihood ratio test has not yet receive much attention in multiple testing procedures. The purpose of this dissertation has been to develop a novel multiple testing procedure based on the likelihood ratio test.

In Chapter 1, I provide a detailed review of several very popular multiple testing procedures. The main focus of Chapter 1 is to show the similarities among those procedures and illustrate that they are all closed test procedures with different choices to test intersection hypotheses. Also included in Chapter 1 is the introduction on partitioning principle which is more advanced than closure principle and has been in the spot light in recent research. At the end of Chapter 1, a natural question is raised as using likelihood ratio test with closure principle or partitioning principle.

In Chapter 2, the difficulty of applying likelihood ratio test with closure principle or partitioning principle is revealed. Several logical consistency requirements are introduced

in this chapter such that a general solution is available to construct consistent multiple testing procedures. This general solution provide us a guideline on how to avoid inconsistency in the closed testing or partition testing settings. Following the guideline, applying likelihood ratio test in those settings is just a special case.

Following Chapter 2, we focus our discussion in Chapter 3 on a very special but popular situation where multiple comparisons with a control problem is of interest. As most of the multiple testing procedures reviewed in Chapter 1 are developed in context of multiple comparisons with a control, it is fairly easy to illustrate the idea of a novel procedure using likelihood ratio test and compare it with existing procedures discussed in Chapter 1. It is the first time in the literature that the closed testing procedure is combined with the likelihood ratio test and a very specific forms of the testing procedure is introduced. Certainly, the key step in the new procedure is to ensure the consistency requirement in multiple testing and a consonance adjustment is introduced. Extensive simulation studies has been provided in this chapter to show the advantage of the new procedure. Also discussed in this chapter is the computation issue for likelihood ratio based test and some remarks are made.

Encouraged by the success of consonance closed likelihood ratio test introduced in Chapter 3, we decided to extend the framework of consonant closed test to other scenarios with greater challenge. For example, the multiple endpoints problem in clinical trials is a very active research topic currently and a powerful multiple testing procedure is demanded by the fast development in pharmaceutical statistics. In Chapter 4, we first try to introduce the multiple endpoints problem and current developments. Then we apply the consonant adjusted likelihood ratio test to the partition testing procedure developed by Liu and Hsu (2009). The new procedure is again very successful as showed in the simulation studies.

Bibliography

Berger, R. L., Hsu, J. C., 1996. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science 11 (4), 283–302.

Bittman, R. M., Romano, J. P., Vallarino, C., Wolf, M., 2009. Optimal testing of multiple hypotheses with common effect direction. Biometrika 96 (2), 399–410.

Bretz, F., Hsu, J., Pinheiro, J., Liu, Y., 2008. Dose finding - a challenge in statistics. Biometrical Journal 50 (4), 480–504.

Bretz, F., Maurer, W., Brannath, W., Posch, M., 2009. A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine 28, 586–604.

Bretz, F., Pinheiro, J., Branson, M., 2005. Combining multiple comparisons and modelling techniques in dose-response studies. Biometrics 61, 738–748.

Chen, X., Luo, X., Capizzi, T., 2005. The application of enhanced parallel gatekeeping strategies. Statistics in Medicine 24, 1385–1397.

Chuang-Stein, C., Stryszak, P., Dmitrienko, A., Offen, W., 2007. Challenge of multiple co-primary endpoints: A new approach. Statistics in Medicine 26, 1181–1192.

Das Gupta, S., Eaton, M., Olkin, I., Perlman, M., Savage, L., Sobel, M., 1972. Inequalities on the probability content of convex regions for elliptically contoured distributions. In: Proceed-

ings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 2. University of California Press, pp. 241–265.

Dmitrienko, A., Fritsch, K., Hsu, J., Ruberg, S., 2007a. Design and analysis of dose-ranging clinical studies. In: Dmitrienko, A., Chuang-Stein, C., D'Agostino, R. (Eds.), Pharmaceutical Statistics Using SAS: A Practical Guide. SAS Press, Ch. 11, pp. 273–311.

Dmitrienko, A., Offen, W., Wang, O., Xiao, D., 2006a. Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. Pharmaceutical Statistics 5, 19–28.

Dmitrienko, A., Offen, W. W., Westfall, P. H., 2003. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Statistics in Medicine 22, 2387–2400.

Dmitrienko, A., Tamhane, A. C., Wang, X., Chen, X., 2006b. Stepwise gatekeeping procedures in clinical trial applications. Biometrical Journal 48 (6), 984–991.

Dmitrienko, A., Wiens, B. L., Tamhane, A. C., Wang, X., 2007b. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. Statistics in Medicine 26, 2465–2478.

Dunnett, C. W., 1955. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association 50 (272), 1096–1121.

Dunnett, C. W., Tamhane, A. C., 1991. Step-down multiple tests for comparing treatment with a control in unbalanced one-way layouts. Statistics in Medicine 11, 1057–1063.

Dunnett, C. W., Tamhane, A. C., March 1992. A step-up multiple test procedure. Journal of the American Statistical Association 87 (417), 162–170.

Dunnett, C. W., Tamhane, A. C., 1993. Power comparisons of some step-up multiple test procedures. Statistics and Probability Letters 16, 55–58.

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

Eaton, M. L., Muirhead, R. J., 2007. On a multiple endpoints testing problem. Journal of Statistical Planning and Inference 137, 3416–3429.

Finner, H., 1994. Testing multiple hypotheses: General theory, specific problems, and relationships to other multiple decision procedures. Habilitationsschrift, FB IV Mathematik, Univ. Trier.

Finner, H., Strassburger, K., 2002. The partitioning principle: A powerful tool in multiple decision theory. The Annals of Statistics 30 (4), 1194–1213.

Gabriel, K., 1969. Simultaneous test procedures - some theory of multiple comparisons. The Annals of Mathematical Statistics 40 (1), 224–250.

Hargé, G., 1998. A correlation inequality for the Gaussian measure. Comptes Rendus de l'Acadmie des Sciences - Series I - Mathematics 326 (11), 1325–1328.

Hochberg, Y., 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics 6, 65–70.

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

Hsu, J. C., 1996. Multiple Comparison: Theory and methods. Chapman & Hall/CRC.

Hsu, J. C., Berger, R. L., 1999. Stepwise confidence intervals without multiplicity adjustment

for dos-response and toxicity studies. Journal of the american statistical association 94 (446), 468–482.

Huang, Y., Hsu, J. C., 2007. Hochberg's step-up method: cutting corners off Holm's step-down method. Biometrika 94 (4), 965–975.

Hung, H. M. J., Wang, S.-J., 2009. Some controversial multiple testing problems in regulatory applications. Journal of Biopharmaceutical Statistics 19, 1–11.

Hyndman, R., Fan, Y., 1996. Sample quantiles in statistical packages. American Statistician 50, 361–365.

Kwong, K., Liu, W., 2000. Calculation of critical values for dunnett and tamhane's step-up multiple test procedure. Statistics and Probability Letters 49 (4), 411–416.

Li, W. V., Shao, Q.-M., 2001. Gaussian processes: Inequalities, small ball probabilities and applications. In: Rao, C., Shanbhag, D. (Eds.), STOCHASTIC PROCESSES: THEORY AND METHODS. Vol. 19 of Handbook of Statistics. Elsevier, Ch. 17, pp. 533–598.

Lin, S., Rogers, J. A., Hsu, J. C., 2001. A confidence-set approach for finding tightly linked genomic regions. American Journal of Human Genetics 68, 1219–1228.

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

Liu, Y., Hsu, J. C., 2009. Testing for efficacy in primary and secondary endpoints by partitioning decision paths. Journal of the American Statistical Association 104 (488), 1661–1670.

Liu, Y., Hsu, J. C., Ruberg, S., 2007. Partition testing in dose-response studies with multiple endpoints. Pharmaceutical Statistics 6, 181–192.

Marcus, R., Peritz, E., Gabriel, K. R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63 (3), 655–660.

Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., Muirhead, R., Stryszak, P., Boddy, A., Chen, K., Copley-Merriman, K., Dere, W., Givens, S., Hall, D., Henry, D., Jackson, J. D., Krishen, A., Liu, T., Ryder, S., Sankoh, A., Wang, J., Yeh, C.-H., 2007. Multiple co-primary endpoints: Medical and statistical solutions. a report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of america. Drug Information Journal 41, 31–46.

Robertson, T., Wright, F. T., Dykstra, R. L., 1988. Order Restricted Statistical Inference. John Wiley & Sons, Inc.

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

Ruberg, S. J., 1995. Dose response studies ii. analysis and interpretation. Journal of Biopharmaceutical Statistics 5, 15–42.

Schechtman, G., Schlumprecht, T., Zinn, J., 1998. On the Gaussian measure of intersection. The Annals of Probability 26 (1), 346–357.

Sonnemann, E., 2008. General solutions to multiple testing problems. Biometrical Journal 50 (5), 641–656, translation with minor corrections of the original article “Sonnemann, E. (1982). Allgemeine Lösungen multipler Testprobleme. EDV in Medizin und Biologie **13**(4), 120–128” by Helmut Finner.

Stefansson, G., Kim, W., Hsu, J., 1988. On confidence sets in multiple comparisons. In: Gupta, S., Berger, J. (Eds.), Statistical Decision Theory and Related Topics IV. Vol. 2. Academic Press, New York, pp. 89–104.

Tamhane, A. C., 1996. Multiple comparisons. In: Ghosh, S., Rao, C. R. (Eds.), Handbook of Statistics: Design and Analysis of Experiments. Vol. 13. Elsevier Science B.V., Ch. 18, pp. 587–630.

Tang, D., Lin, S., 1997. An approximate likelihood ratio test for comparing several treatments to a control. Journal of the American Statistical Association 92, 1155–1162.

Tang, D.-I., Gnecco, C., Geller, N. L., 1989. An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. Biometrika 76 (3), 577–583.

Wassmera, G., Reitmeir, P., Kieser, M., Lehmacher, W., 1999. Procedures for testing multiple endpoints in clinical trials: An overview. Journal of Statistical Planning and Inference 82, 69–81.

Westfall, P., Tobias, R., Rom, D., Wolfinger, R., Hochberg, Y., 1999. Multiple Comparisons and Multiple Tests: Using the SAS System. SAS Institute, Cary, NC.

Westfall, P., Young, S., 1993. Resampling-based multiple testing. John Wiley & Sons, Inc., New York.

Xu, H., Nuamah, I., Liu, J., Lim, P., Sampson, A., 2009. A Dunnett-Bonferroni-based parallel gatekeeping procedure for dosecresponse clinical trials with multiple endpoints. Pharmaceutical Statistics 8, 301–316.