Model Selection Procedure with Familywise Error Rate Control for Binomial Order-Restricted Problems

Von der Naturwissenschaftlichen Fakultät
der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des
akademischen Grades eines

Doktors der Gartenbauwissenschaften
- Dr. rer. hort. -

genehmigte Dissertation

von

M.Sc. **Xuefei Mi**
geboren am 13.05.1980 in Shandong, P.R.China

Hannover, 2009

Referent: Prof. Dr. Ludwig Hothorn

Korreferent: Prof. Dr. Walter Lehmacher

Tag der Promotion: 23.03.2009

Zusammenfassung

Modell-basierte und Test-basierte Methoden werden meistens zur Analyse der Dosis-Wirkung Beziehung verwendet. Der Gebrauch von ordnungseingeschränkten Hypothesen ist eine einheitliche Methode, die Power vergrößert, den alternativen Raum einengend. Hiermit sind Änderungspunkte, einfache Ordnung und einfacher Baum drei allgemeine Typen von Ordnungsbeschränkungen.

Wir werden diese zwei Methoden im Änderungspunkte-Entdeckungsproblem und anderen Ordnungsbeschränkungsproblem vergleichen. Die Modell-gegründete Methode konzentriert sich auf, wie man die reale Information herausfindet. Auf der anderen Seite konzentriert sich die Test-gegründete Methode auf, wie man eine Verteilung aufbaut und das FWER kontrolliert. Nach dem Vergleich stellen, auch werden wir ein modifiziertes Informationskriterium präsentieren, das die FWER für die Musterauswahl unter der bestimmten Ordnungsbeschränkung kontrollieren kann.

Schlagworte: multiple Kontrasttests, ordnungseingeschränkten Hypothesen, Informationskriterium

Abstract

Model-based (Royston et al., 1999) and test-based (Dosemeci and Benichou, 1998) methods are most commonly used to analyze dose-response relationship. The use of order-restricted hypotheses is a common approach which increases power by narrowing down the alternative space. Hereby, Change-point, Simple-order and Simple-tree are three common types of order restrictions.

In this thesis, we will compare model-based methods with test-based methods in Change-point detection and other order restriction problems. The model-based method focuses on the real information distance. On the other side, the test-based method focuses on how to build a distribution and to control the Familywise error rate (FWER). After the comparison, we will also present a new method, called Multiple Likelihood Test (MLT), which can control the FWER for model selection under different order restrictions. First, we build Mi and Hothorn Information Criterion (MHIC) to do model selection. We will consider the null hypothesis and all the elementary alternative models as candidate models. Second, the information differences between the null model and the elementary alternative models will be calculated. Finally, we will build the distribution of these differences and calculate the critical value to control the FWER. In order to solve the "overfitting" problem, we also modify the maximum likelihood estimators (MLE) into suitable likelihood estimators (SLE) for calculating the information criterion under certain order restrictions, such as Simple-order restriction and Simple-tree oder restriction.

Keyword: Multiple Contrast Test, Order-restricted hypothesis test, Information Criterion

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is focused on the application of statistics the fields in biology, epidemiology and pharmacology. The mount of data in this area is huge and there are always random effects within the data. Therefore, we usually do not count the original figures one by one. The data will be arranged in several groups. Appropriate models are built to describe the relationship among these groups.

For example, "Exposed with different levels of sulfidic nickel, are nickel-refinery workers more likely to get lung cancer than others?", "How do the ages of the parents effect the spontaneous abortion rate?" and "Is there any structure change between functional motifs and 'junk-DNA' on the genomes?" are frequently asked questions for building the appropriate models.

In the questions mentioned above, the null hypothesis, which usually assumes no differences or no changes, should be tested first. After the null hypothesis is rejected at a certain global level α, a model selection procedure is used to detect the relationship. However, the test and model selection procedure after the test depend seriously on the dose-response shape, which is unknown before the whole procedure. In this situation, a procedure, which has good power over the order-restricted alternative space, is required.

We will compare model-based method with test-based method in solving order restricted problems. The previous knowledge of order, such as "The effects of higher

dose are stronger than lower dose and placebo" (Simple-order restriction) give more power to the test.

An order restricted likelihood ratio test, which is test-based method, was developed by Robertson et al. (1988). Chaudhuri and Perlman (2005) gave a former mathematical definition to the common order restriction of Simple-order and Simple-tree order restriction, which will be discussed in Chapter 3. The idea of taking advantage of order restriction is widely used in testing methods. Several test-based approaches are available for these problems, such as max-t statistics (Hirotsu and Srivastava, 2000), which is test-based and is formulated as maximum contrast approach belonging to the broader class of Multiple Contrast tests (MCT). Mukerjee et al. (1987) also recommended orthogonal contrasts which have a simple power function .

The Multiple Contrast test (MCT) uses the Maximum Likelihood Estimators (MLE) to reject the global hypothesis and select the model with the largest test statistics as the most possible model. Chaudhuri and Perlman (2005) also proved that this estimator achieves the smallest squared error. However, the estimators which have largest test statistics are just the value which fit the data best. The disadvantage of MCT is that the score functions of MCT are not designed to do model selection, although we can select model via them. It does model selection if and only if the global null hypothesis is rejected. The local finding rate of the true model of the alternative is low.

Instead of using MCT, Bretz et al. (2005) suggested using Information Criterion (IC) which is model-based method and is designed to do model selection. The famous Akaike Information Criterion (AIC) uses Kullback-Leibler distance to punish models with numbers of unknown parameters (Akaike, 1974). Following the similar idea, Anraku (1999) invented Order Restricted Information Criterion (ORIC) for normal distributed data by using the same Kullback-Leibler distance to achieve the best estimator which has the largest adjusted log-likelihood under simple-order restriction. An algorithm for binomial data is also available (Hothorn et al., 2008). Ninomiya (2005) also gave the penalty term under change-point order restriction. Zhao and Peng (2002) developed a penalty term which is proportional to the loga-

rithm of the total sample size. Most of these methods focused on how to estimate the adjusted unbiased information of different elementary hypothesis. The complicated distribution of the Information Criterion increases the difficulty of building a test. "Furthermore, when the number of hypotheses increases, the adjustment of multiplicity should also be considered." Robertson et al. (1988) proved that the value of the log-likelihood ratio between the null hypothesis and the alternatives is weighted chi-squared distributed under simple-order restriction. In another point of view, Vuong (1989) focused on the distribution of the log-likelihood ratio under different model structures. The distributions of the log-likelihood ratio for non-nested, nested and overlapping models are developed by him. Xiong and El Barmi (2002) developed a non-parametric penalty term which selects the correct model among multiple hypotheses with control of FWER.

These model-based methods use likelihood to find the most possible model. By using pool-adjacent-violators algorithm (PAVA) (Robertson et al., 1988), the global maximum likelihood estimator (gMLE) can be achieved. By using the gMLE, these methods have better finding rate of the true model, but have no control of FWER. They treat the null model as one of the possible models among all others. These methods are not constructed as hypothesis tests to reject the alternatives.

In this thesis, we will present a Multiple Likelihood Test (MLT), which is test-based model selection method. It can control the FWER for model selection under different order restrictions. Our research is based on AIC, which has the idea of combining information and likelihood. Our method is also similar to Anraku (1999); Xiong and El Barmi (2002). But the difference between them can be described as follows: Anraku (1999) looked for bias-adjusted information criterion under Simple-order restriction. In general, he focused on the mean of the information. On the other hand, Xiong and El Barmi (2002) were trying to do model selection with controlling the FWER under the null hypothesis, and they simulated the critical value. In this thesis, we extend the work of Robertson et al. (1988) to generate a parametric penalty term for model selection under different order restrictions with control of FWER. Examples are given in Chapter 2. In Chapter 3, we define and build the model,

review the previous methods and give solutions for every order restrictions. Chapter 4 includes, the general theory to calculate the Information Distance and the penalty term. In Chapter 5, we introduce our new test-based model selection method: Multiple Log-likelihood Test (MLT). The relationship between model selection and test is presented too. Power study will be given in Chapter 6. We also conduct a simulation study to evaluate the new parametric penalty term and compare the correct model selection rate with the previous methods. Chapter 7 provides software to calculate multivariate normal distribution and to make the model selection with control of FWER. Discussion and summary are given in Chapter 8.

Chapter 2

Motivations

In this chapter, examples of binomial order-restricted problems are presented. Detailed discussion will be given in the following chapters.

2.1 Dose-Response Studies

2.1.1 A clinical dose finding study with an adverse events rate

In pharmacology, drugs are the chemical substances which are used to prevent, treat or cure disease. However, the drugs are commonly associated with adverse events, if the patients overdose. The following example in Table 2.1 is part of a clinical dose finding study with adverse events rate (Bretz and Hothorn, 2002). Placebo or cabergoline in different dosages are given to the patients twice a week. Adverse events are observed in both of the placebo group and dosage groups. The researchers want to know if the adverse rate increases markedly at certain dose level of cabergoline. If the answer is yes, how much is this dose level?

This is a special case of Single Change-point order restriction (definition will be given in Chapter 3.1.1). From the given data, we could guess that one Change-point exists between the lower dose (0.125mg) group and the higher dose (1.0mg) group.

Treatment	Placebo	0.125(mg)	1.0(mg)
Present x_i	9	19	24
Absent $n_i - x_i$	11	24	17
Total n_i	20	43	41
$\hat{p}_i$	0.45	0.44	0.58

Table 2.1: Presence or absence of adverse events.

Exposure	*Unexp.*	*Low*	*LtoM*	*Medium*	*MtoH*	*High*
Cancer x_i	10	27	48	42	40	46
Normal $n_i - x_i$	57	93	95	92	94	94
Total n_i	67	120	143	134	134	140
$\hat{p}_i$	0.149	0.225	0.336	0.313	0.299	0.329

Table 2.2: Lung cancer and exposure to nickel.

5

2.2 Epidemiological case-control studies

2.2.1 Sulfidic nickel and lung cancer

Nickel compounds are classified as carcinogenic to humans by the International Agency for Research on Cancer. The next example is part of a case-control study of Norwegian nickel-refinery workers (Grimsrud et al., 2002). The total amount and quantification of sulfidic nickel in the working area for different workers are recorded. The researchers want to know if there is a clear dose-dependent increase in their studies. If there is a trend, of what type is it?

First they want to test if the Nickel compounds induce cancer. Then, a suitable dose-response model for the effect of sulfidic nickel in lung cancer should be selected. This problem can be interpreted as model selection under Simple-order restriction (definition will be given in Chapter 3.1.3).

2.2.2 The effect of age on the spontaneous abortion rate

In general knowledge, women might have a higher spontaneous abortion rate if they become pregnant at a higher age. However, this general knowledge is only "partly

Males age	< 25	$25 - 29$	$30 - 34$	$35 - 39$
Abortion x_i	33	37	3	7
Normal $n_i - x_i$	226	321	61	5
Total n_i	259	358	64	12
$\hat{p}_i$	0.127	0.103	0.047	0.583

Table 2.3: Spontaneous abortion rate and age of the father.

true". The rate also depends on the age of the father. Slama et al. (2003) investigated the percentage rate of spontaneous abortion between weeks 5 and 20 of pregnancy according to the age of the parents. A random cross-sectional population of 1,151 French women, who had been pregnant between 1985 and 2000, are interviewed by telephone. The strata of 20-24 years old females is selected and presented in Table 2.3.

In this example, the researchers are interested in how does the age of the father effect the abortion rate. They not only want to reject the null hypothesis, but also want to know which group has higher abortion rate. If one of the parents' age is fixed, such as the data in our table, a model selection procedure under Simple-tree order restriction for Many-to-one comparison or under Simple-order restriction for trend test, is suitable for this case (definition will be given in Chapter 3.1.3 and 3.1.4).

2.3 Bio-informatics: DNA-motif finding

Huge amount of data is given in the research field of bio-informatics which is a crossover science between mathematics and biology. The problem is data mining rather than testing. Interesting candidates with some special genetic meanings should be found, as much as possible. Biologists use these information in next stage to uncover further relationship. A false positive report of candidate could waste their money and time in building experiments. However, researchers care less about the control of FWER than the motif finding.

In DNA sequences, different positions have different degrees of conservation. The DNA-binding proteins are bound to some very conservative base pairs, called motif

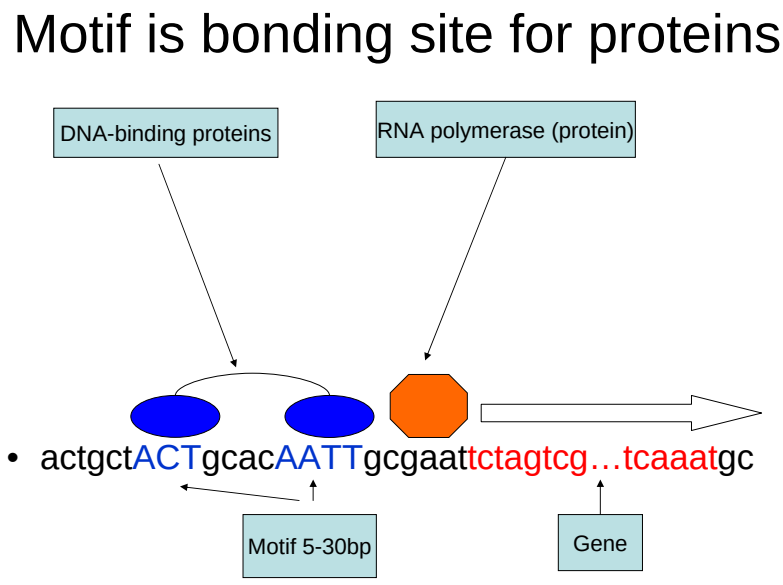

Figure 2.1: Property about DNA Motif (Lewin, 2004).

(Lawrence and Reilly, 1990). The motif, which usually has a length from 5 bp (base pairs) to 30 bp, is followed by one or more genes. In the protein producing procedure, first a DNA-binding protein bounds itself to the motif, then RNA polymerase will follow the binding protein and decode the gene region for producing new functional proteins. The procedure of DNA-binding is shown in Figure 2.1.

Usually, the motif finding procedures are carried out by two steps, alignment and comparison. Alignment is a kind of algorithm, which finds candidate regions where the motif might locate. Many papers have discussed about how to align the motifs for finding the possible regions, such as the L-tuple method. The L-tuple method cut the DNA sequences into small tuples. By prior assumption, the motifs exist in every sequences. The frequency of the L-tuples, which are part of the motifs, are higher in every sequences. L-tuple method picks up the higher frequency tuple and aligns them into the DNA sequences to recover the motifs (Pevzner et al., 2001).

After alignment, the candidate regions are put together for detail comparison to find out the exact code of the motif. Fourteen instances of the Gal4 binding site motif

*atactt*CGGAGCACTGTTGAGCG	(2.1)
*agcgct*CGGACAACTGTTGACCG	(2.2)
CGGCGGCTTCTAATCCG	(2.3)
*t*CGGAGGGCTGTCGCCCG	(2.4)
CGGAGGAGAGTCTTCCG	(2.5)
*attgtt*CGGAGCAGTGCGGCGCG	(2.6)
CGCGCCGCACTGCTCCG*aacaat*	(2.7)
CGGAAGACTCTCCTCCG	(2.8)
CGGGCGACAGCCCTCCG*a*	(2.9)
CGGATTAGAAGCCGCCG	(2.10)
*tat*CGGGGCGGATCACTCCG*aac*	(2.11)
*cac*CGGCGGTCTTTCGTCCG*tgc*	(2.12)
CGGCGCACTCTCGCCCG	(2.13)
*t*CGGGGCAGACTATTCCG*g*	(2.14)

Table 2.4: Aligned DNA motif: Saccharomyces Cerevisiae Promoter Database (SCPD) (Zhu and Zhang, 1999)

in yeast, from the Saccharomyces Cerevisiae Promoter Database (SCPD) (Zhu and Zhang, 1999), are listed in Table 2.4. Characters with upper case are the aligned part. After alignment, the motifs are put together to get compared. Some parts are very conservative while some other parts are less conservative or like random. For example, In the motif matrix, the first column of the aligned motif has 14 "C", the second column has 13 "G", only one "C". In these situations, the first and second columns are conservative parts. Similarly, column 3, 15, 16 and 17 are conservative parts too. The columns, which are not mentioned above, are the less conservative part. For example, the eighth column has eight "C", one "T" and five "C".

The matrix Y = $\{y_{i,j}$, i=1,2,3,4; j=1,...,k$\}$ in Table 2.5 is estimated by counting the numbers of the bases A, C, G and T along the columns of Table 2.4. The $\{A, C, T, G\}$ are assumed as multinomial distribution.

A list of sequences of length k is often described by a 4 by k Position specific Weight Matrix (PWM), $\theta = \{\theta_{i,j}$, i=1,2,3,4; j=1,...,k$\}$ (Stormo et al., 1982). The PWM are widely used in motif finding for simplicity in modeling and calculating. The PWM

	$y_{i,j}$	j=1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
No. of A	i=1	0	0	0	7	1	1	9	0	6	1	0	3	1	2	0	0	0
No. of C	i=2	14	0	1	3	3	6	3	8	0	5	3	7	5	3	12	14	0
No. of G	i=3	0	14	13	4	9	6	1	5	0	6	1	2	6	1	2	0	14
No. of T	i=4	0	0	0	0	1	1	1	1	8	2	10	2	2	8	0	0	0

Table 2.5: 4 by k frequency table.

θ, which is gotten from the 4 by k frequency table, is given as following

$$\theta = \frac{1}{14} \times \begin{Bmatrix} 0 & 0 & 0 & 7 & 1 & 1 & 9 & 0 & 6 & 1 & 0 & 3 & 1 & 2 & 0 & 0 & 0 \\ 14 & 0 & 1 & 3 & 3 & 6 & 3 & 8 & 0 & 5 & 3 & 7 & 5 & 3 & 12 & 14 & 0 \\ 0 & 14 & 13 & 4 & 9 & 6 & 1 & 5 & 0 & 6 & 1 & 2 & 6 & 1 & 2 & 0 & 14 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 8 & 2 & 10 & 2 & 2 & 8 & 0 & 0 & 0 \end{Bmatrix} \tag{2.15}$$

In the conservative part (column 1, 2, 3, 15, 16 and 17), the frequency of some main bases is very high. In the meanwhile, the frequencies of some other bases are very low, we can simply assume that the other three bases are random error with equal probability. In the less conservative part (e.g. position 4), some base (e.g. bases A in position 4) has little higher occurrence rate than the other three, which are usually not random errors. For simplicity, we just assume that the four bases in less conservative parts are random error.

According to van Zwet et al. (2005), the entropies are calculated over all bases

$$H_j(\theta) = -\Sigma_{i=1}^{4} \theta_{i,j} \log \theta_{i,j},\ j = 1, ..., 17 \tag{2.16}$$

in here, they consider the error as random error too.

Let us assume that the distribution of the dominate base, which is the maximum in each column, is binomial and the distribution of the sum of the other less conservative base is binomial too. The bases which have the largest occurrence are taken as the most possible bases and are assumed to be binomial distributed after taking maximum over the data. A contingency table is made in Table 2.6. In the table, x_i

Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
x_i	14	14	13	7	9	6	9	8	8	6	10	7	6	8	12	14	14
n_i	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
p_i	1	1	.9	.5	.6	.4	.6	.6	.6	.4	.7	.5	.4	.6	.9	1	1

Table 2.6: Contingency table

are achieved by taking the maximum of the i-th column in the frequency Table 2.5, i.e. $x_j = \max_i y_{i,j}$, $n_j = \sum_i y_{i,j}$, $p_j = \frac{x_j}{n_j}$

In the position 4 and 14 of this table, two change points with the differences around 0.4 are suspected. The motif appears to have a high-low-high structure, which is defined as Epidemic-order restriction (definition will be given in Chapter 3.1.2), which is one common structure of a motif.

$$\overbrace{p_0 = ... = p_{s-1}}^{High\,part} > \underbrace{p_s = ... = p_{j-1}}_{Low\,part} < \overbrace{p_j = ... = p_k}^{High\,part}, 0 < s < j < k \tag{2.17}$$

The high part is the epidemic state. Length of low part is larger than 3. It is the less conservative part. Length of high parts is larger than 2. They are very conservative parts. In the last example sequence length $k = 17$, the pattern of high-low-high is 3-11-3.

The starting point of this thesis is the order restriction model according to van Zwet et al. (2005). Their approach used the Position-specific Weight Matrix for modeling the motif. Then, they used the regression model based on MLE under specific order restriction. However, no parameter adjustment is applied to these regression models. They implemented three versions of the algorithm for unaligned data.

An improved method is demonstrated in this thesis. The same PWM are transformed into binomial model. The values in the very conservative part are pooled to take the average. The advantage of this method is that it uses the dependent information of the conservative motif and uses parameter adjustment to have a better selection result.

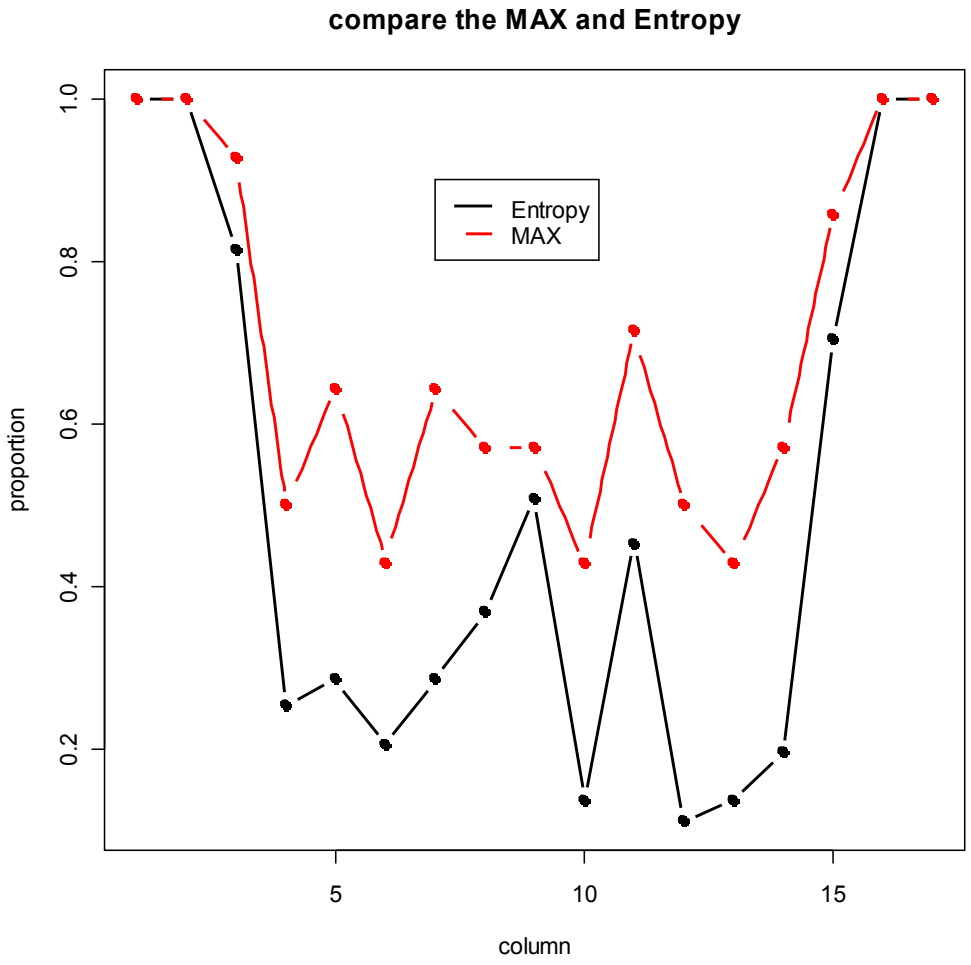

Figure 2.2: The maximum and entropy values (van Zwet et al., 2005) of the aligned DNA motif

Chapter 3

Model selection procedure

In this chapter, we will first build the mathematical model for the binomial order-restricted problems given in previous chapter. Secondly, we will give a general introduction of model selection. Finally, we will give a brief review of the previous test methods and model selection methods for these problems.

3.1 Order restriction

3.1.1 Single Change-point order restriction

Let random variables $X_0, X_1, ..., X_k$ be binomial distributed with sample size n_i and proportion $p_i = x_i/n_i$, $i = 0, ..., k$ where the observations x_i are generated from the distribution of X_i. The hypotheses can be formulated as

$$\begin{aligned} H_0 &: p_0 = p_1 = ... = p_k \\ H_A &: \bigcup_{j=1}^{k} H_A^j \end{aligned} \tag{3.1}$$

here, $j-1$ to j is the position of Change-point. We want to reject H_0 against H_A with controlling the FWER over all k Single Change-points. When the H_0 is rejected with control of FWER, we want to select the elementary model, which has the largest

Hypothesis	$\tilde{p_0}$	$\tilde{p_1}$	$\tilde{p_2}$
$H_0 : p_0 = p_1 = p_2$	0.5000	0.5000	0.5000
$H_A^1 : p_0 < p_1 = p_2$	0.4500	0.5119	0.5119
H_A^2: $p_0 = p_1 < p_2$	0.4444	0.4444	0.5854

Table 3.1: Region of different hypothesis of the adverse events rate.

test statistics, as the best model. The global alternative can be decomposed into k elementary Single Change-points (Hirotsu and Marumo, 2002).

$$H_A^j : p_0 = ... = p_{j-1} < p_j = ... = p_k,\ j = 1, ..., k \tag{3.2}$$

The dose finding study can be solved as Single Change-point detection problem, if the researchers want to know on which dose the reverse effect increase significantly. The hypotheses are listed in Table 3.1. The region of different hypotheses can also be seen in Picture 3.1. In this picture, each point in the three dimension space represents the three proportions for the three dose levels. x is the data we observed. H_0, H_A^1, H_A^2 are the candidate models we want to select. $H0, H1, H2$ are the estimated parameters in different models i.e $H0 \in H_0$, $H1 \in H_A^1$, and $H_2 \in H_A^2$. We simulate the random data points from these three different estimators. Spheres with different colors are the 95% confidence regions for corresponds models. From the picture we see that X is close to point $H2$, so model H_A^2 is the "best" candidate to be selected. However, these three spheres are overlapped and X is inside the overlapping region. If we select H_A^2 as the best model, we cannot control the error rate. So here we suggest the researcher increase the sample size to achieve higher power and lower error for the selection.

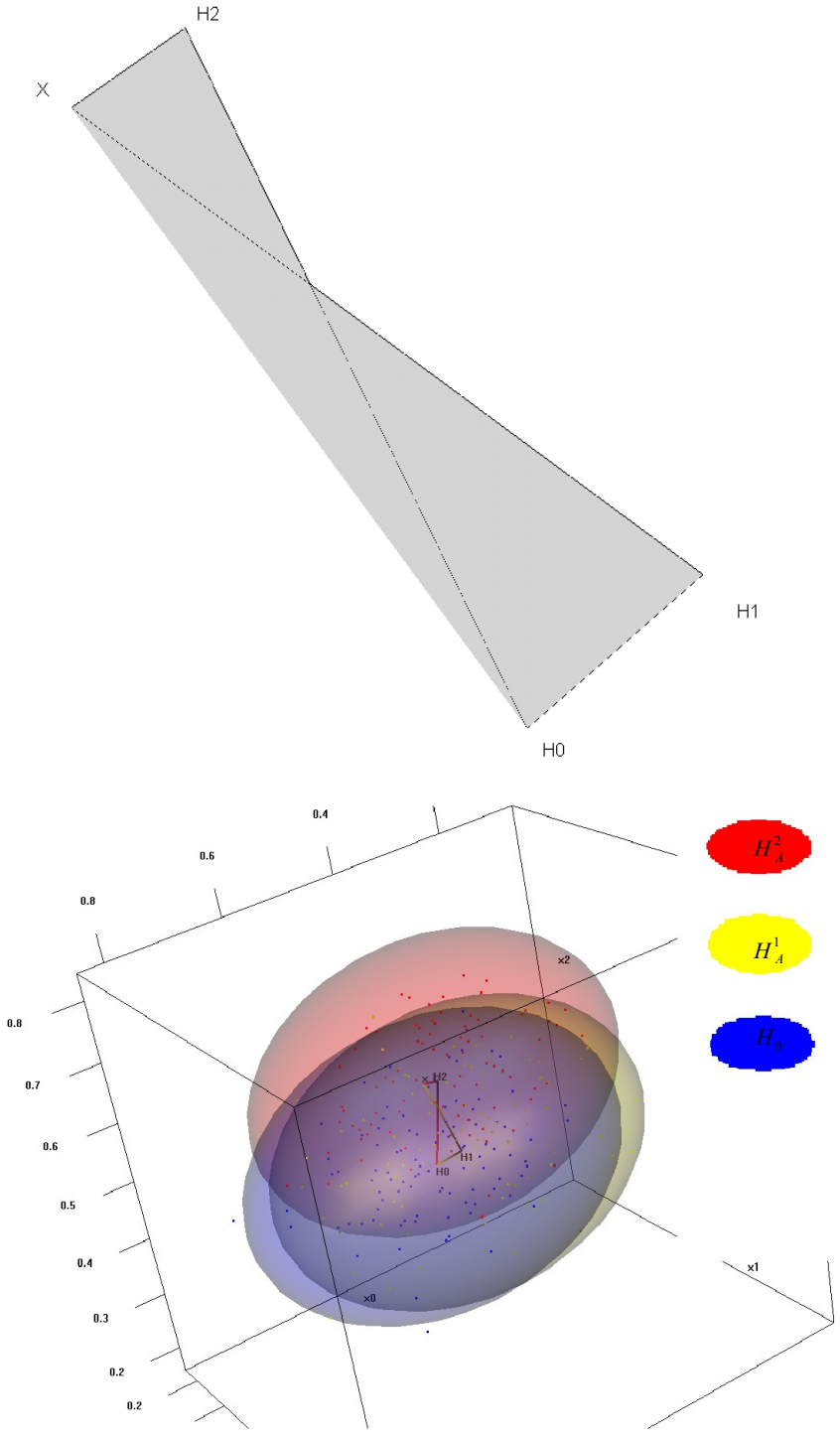

Figure 3.1: Simulation for the data: Points are generated proportions; Spheres are the 95% confidence regions

3.1.2 Epidemic-order restriction

The Epidemic hypotheses can be formulated as

$$H_0 : p_0 = p_1 = ... = p_k$$
$$H_A : \bigcup_{s=1,j=2}^{k,k} H_A^{s,j} \tag{3.3}$$

here, s, j are the positions of the down and up Change-points. The possible combination number of these situations is calculated from the combination rules. We should select two positions out of k positions, i.e. $k!/2!(k-2)! = k(k-1)/2$. We want to reject H_0 against H_A globally with controlling the FWER over all these $k(k-1)/2$ alternatives. When the null is rejected with control of FWER, we want to select one of the elementary model, which has the largest test statistics, as the best model. The global alternative can be decomposed into $k(k-1)/2$ elementary ones.

$$H_A^{s,j} : p_0 = ... = p_{s-1} > p_s = ... = p_{j-1} < p_j = ... = p_k,\ 0 < s < j < k \tag{3.4}$$

Epidemic-order restriction can be considered as a special restriction which has two Single Change-points. So we give the reference of Single Change-point and Epidemic-order restriction together. Yao (1993) developed a test for normal data under Epidemic-order restriction. Ninomiya (2005, 2006) developed an Information Criterion (IC) for normal and binomial data under Epidemic-order restriction. As we have already mentioned in our previous example, the DNA-motif finding problem can be considered as a model selection procedure for binomial or multinomial data under Epidemic-order restriction. A test for binomial data under Change-point order restriction and its application in sequence analysis of HIV are given by Halpern (1999):" accidental recombination between two viral genomes present in the same host cell, the genetic sequence of the offspring involves adjacent regions derived from the two parental sequences, with a sharp Change-point in between."

3.1.3 Simple-order restriction

The hypotheses can be described as

$$\begin{aligned} H_0 &: p_0 = p_1 = ... = p_k \\ H_A &: \bigcup_{j=1}^{t} H_A^j \end{aligned} \tag{3.5}$$

here, H_A should have at least one inequality strict. We want to reject H_0 against H_A globally with controlling the FWER over all k alternatives. When the null is rejected at global level α, we want to find out all of the inequalities. Simple-order restriction can be considered as a special restriction which has many Single Change-points. The global alternative can be decomposed into $2^k - 1$ elementary ones (Hothorn et al., 2008; Hirotsu and Marumo, 2002; Bretz and Hothorn, 2002; Robertson et al., 1988). Here we use i to note the number of Single Change-points. Let C be the combination calculator. There are in total C_k^i kind of combination, if there is exactly i in the model Single Change-points. We use j, $1 \leq j \leq C_k^i$ to note the j-th combination among those.

The elementary model for the j-th model with exactly i Single Change-points is

$$H_A^j : \underbrace{p_0 \leq ... \leq ... \leq p_k}_{i \quad inequalities} \tag{3.6}$$

here, $t = j - 1 + \sum_{l=1}^{i-1} C_k^l$ is the order of this model over the whole alternative models.

Hirotsu and Marumo (2002) introduced a MCT method to make the multiple decision under Simple-order restriction. Xiong and El Barmi (2002) also developed an Information Criterion for this problem. Many application are founded for such type of order restriction. For example, the spontaneous abortion rate study, introduced in last chapter, can be solved as Simple-order problem. The researchers want to know if the age of the fathers effects the rate. The hypotheses are listed in Table 3.2.

Hypothesis	$\tilde{p}_{j,0}$	$\tilde{p}_{j,1}$	$\tilde{p}_{j,2}$	$\tilde{p}_{j,3}$
H_0: $p_0 = p_1 = p_2 = p_3$	0.115	0.115	0.115	0.115
H_A^1: $p_0 < p_1 = p_2 = p_3$	0.115	0.115	0.115	0.115
H_A^2: $p_0 = p_1 < p_2 = p_3$	0.113	0.113	0.131	0.131
H_A^3: $p_0 = p_1 = p_2 < p_3$	0.107	0.107	0.107	0.583
H_A^4 : $p_0 < p_1 < p_2 < p_3$	0.107	0.107	0.107	0.583
H_A^5 : $p_0 = p_1 < p_2 < p_3$	0.107	0.107	0.107	0.583
H_A^6 : $p_0 < p_1 = p_2 < p_3$	0.107	0.107	0.107	0.583
H_A^7 : $p_0 < p_1 < p_2 = p_3$	0.113	0.113	0.131	0.131

Table 3.2: Region of different hypothesis of the spontaneous abortion rate. $\tilde{p}_{j,i}$ are the MLE

3.1.4 Simple-tree restriction

The hypotheses can be described as

$$\begin{aligned} H_0 &: p_0 = p_1 = ... = p_k \\ H_A &: \bigcup_{j=1}^{k} H_A^j \end{aligned} \tag{3.7}$$

here, j is the group which is higher than control. We want to reject H_0 against H_A globally with controlling the FWER over all k alternatives. When the null is rejected at global level α, we want to select all of the elementary models, which are larger than the control. The global alternative can be decomposed into k elementary ones (Robertson et al., 1988).

$$H_A^j : p_0 < p_j \tag{3.8}$$

Simple-tree order restriction can also be interpreted as Many-to-one comparison, since many treatment group are compared to one control group. Dunnett (1955) introduced the MCT method for this problem. Chaudhuri and Perlman (2005) analyzed the mean squared error for the one treatment and one control group. There is no further IC method for this problem. In this thesis we are trying to use Mi and Hothorn Information Criterion (MHIC) with Suitable Likelihood Estimator (SLE) to solve it.

Hypothesis	$\tilde{p_0}$	$\tilde{p_1}$	$\tilde{p_2}$
H_0: $p_0 = p_1 = p_2$	0.5000	0.5000	0.5000
H_A^1: $p_0 < p_1 = p_2$	0.4500	0.5119	0.5119
H_A^2: $p_0 = p_1 < p_2$	0.4444	0.4444	0.5854

Table 3.3: Region of different hypothesis of the adverse events rate. $\tilde{p}_i$ are the MLE

Many application are founded for such type of order restriction. For example,the dose finding study can be solved as Simple-tree problem. The researchers want to know on which dose the reverse effect is significantly different to the control. The hypotheses are listed in Table 3.3. Schaarschmidt et al. (2009) also introduced simultaneous confidence intervals for this Many-to-one comparisons.

3.2 Definition of model selection

"Statistical modeling is a crucial issue in scientific data analysis. Models are used to represent stochastic structures, predict future behavior, and extract useful information from data." (Konishi and Kitagawa, 2008). A good statistical model extracts useful information from observed data. Then it uses this information to represent the stochastic structures. Finally, it uses the structure to predict the future outcomes. The last step is more important than any others in real applications. A parametric model H is decided by finite or infinite many of parameters θ.

"Model selection is the task of selecting a statistical model from a set of potential models, given data" (Burnham and Anderson, 2002). Generally, the number of possible models is infinite. The researchers must define a finite set of models. The true model is assumed to be "well" described by one model from this finite model set. Once the sets are decided, the statistical methods allow us to select one model from these as the best model. The word "well" has many definitions, which focus on different principles, such as "unbiased estimator", "section average", "least square error", "maximum likelihood", "smallest information distance" and "smallest future product error". Here we use T to note the function, which can calculate this predefined standard.

The models we used here are sub sets of the parameter space. We assume that the observed data vector is $\mathbf{x} = \{x_1, x_2, ..., x_k\}$, which is generated from the true parametric model H. We have candidate model $H_0, H_1, ..., H_q$, $(q < k)$ and want to select one of them as an estimation of the best model. The value of $\mathbf{x}$ is projected into each model to find out the most likely parameters θ which generates $\mathbf{x}$. $\hat{\theta}_j(\mathbf{x}) \in H_i$, is the projection point of $\mathbf{x}$ on the space of H_i. We select the model H_j, which has the largest "score" to "well" describe the true model, as the best estimation of the true model H. The value of j can be estimated as follow

$$j = \arg\max_i \{T(\hat{\theta}_i(\mathbf{x})|\mathbf{x}, \hat{\theta}_i(\mathbf{x}) \in H_i\} \tag{3.9}$$

the maximum of the score is noted as

$$T_{max} = \max_i \{T(\hat{\theta}_i(\mathbf{x})|\mathbf{x}, \hat{\theta}_i(\mathbf{x}) \in H_i\} = T_j \tag{3.10}$$

Here $T(\theta)$ is the score function. There are many criterion with good interpretation that can be chosen as the score function. E.g likelihood ratio or weighted square distance which will be discussed in detail in the next following sections.

A good model selection method should achieve the balance between the "goodness of fit" and complexity, in order to give a accurate future prediction. The more complex the models are, the better the models fit the data. However, the observed evidences are just part of whole event space. Since truth can never be uncovered, the complex models just fit the observed data, but not the true model. The "goodness of fit" can be described by maximum likelihood and the complexity increases if the number of unknown parameter increases. Akaike (1974) found a connection between them and introduced Akaike Information Criterion (AIC) as a model selection principle. He defined the good model as the one that has the closest distribution distance to the true model. As mentioned in the examples, the researchers not only want to do model selection from possible models, but also want to test all of these elementary models against the null model. In the following section, we give the definition of

familywise error rate, which controls the error for this multiple pairwise test.

3.3 Familywise error rate (FWER) control

In hypothesis test, usually the alternative hypothesis H_A will be tested against the null hypothesis H_0, which is usually the simplest hypothesis that we have mentioned in former sections. The score function T_0 of the H_0 is also easy to find. In hypothesis test, it is usually the critical value, $T_0 = z_\alpha$.

We use the value of the score function from alternative model to accept or reject the hypothesis. When the maximum value of the score function $T(\theta)$ is larger than certain critical value z_α, we can reject the null hypothesis. However, it might happen that the true model is H_0 and the highest score is still larger than the critical value. In this situation we will reject the null hypothesis by mistake, when it is still true! This event is noted as

$$\{T > z_\alpha | H_0\} \tag{3.11}$$

In this situation, a Type I error (also noted as false positive) is made by the test. In general, this error will be reduced by simply increasing the value of critical value.

When we perform a multiple pairwise test, the familywise error rate (FWER) can be defined as a probability that the number of false discoveries of this test is not zero, i.e.

$$FWER = 1 - P(V = 0) = P(V \geq 1) \tag{3.12}$$

where, V is the number of false discoveries (Type I error). The event of no false discovery is equivalent to the event that the test statistics T_{max} for the multiple

testing procedure is not significant when the null hypothesis is true,

$$\{V = 0\} = \{T_{max} \leq z_\alpha | H_0\} \tag{3.13}$$

As discussed in the last section, we use the score function $T(\theta)$ to value every model. When the true model is H_0 i.e. $\theta \in H_0$, the null model H_0 should be selected as the best model. However, we could still select some other models, which have the largest "score" to "well" describe the true model, as the best estimation of the true model. In this case we declare a false positive.

The probability to make such error is

$$\begin{aligned} FWER &= P(V \geq 1) \\ &= P(T_{max} > z_\alpha) \\ &= P(\max_i \{T(\hat{\theta}_i(X)|X), \hat{\theta}_i \in H_i\} > z_\alpha \} | \theta \in H_0) \end{aligned} \tag{3.14}$$

which is equivalent to

$$FWER = 1 - P(\max_i \{T(\hat{\theta}_i(X)|X), \hat{\theta}_i \in H_i\} \leq z_\alpha \} | \theta \in H_0) \tag{3.15}$$

3.4 Previous test methods

In this section, we will review three test methods which focus on error rate control. They are Likelihood Ratio Test (LRT) (Bartholomew, 1959), Multiple Contrast Test (MCT) (Robertson et al., 1988) and Cochran-Armitage Test (CAT) (Cochran, 1954; Armitage, 1955).

Although LRT was introduced for problems under Simple-order, LRT for problems under Simple-tree exists too. The LRT uses the ordered structure of the mean to gain good average power over the whole alternative space. In general, the power can be optimized via maximizing the likelihood ratio. However the exact likelihood function is complicated to calculate. The power function is even harder to achieve,

so MCT, which has similar behaviors as LRT and is easier to calculate than LRT, is invented. The LRT can be represented or approximated by a correspond MCT.

In contradict to achieve good global power, CAT is a locally most powerful test to detect certain trend. This property of CAT is useful, when some extra information of the model is known. E.g. Additive effect is often assumed in a whole genome association study.

3.4.1 Likelihood Ratio Test (LRT)

Let random variables $X = \{X_0, X_1, ..., X_k\}$ be binomial distributed with proportion vector $P = \{p_0, p_1, ..., p_k\}$ and sample size $N = \{n_0, n_1, ..., n_i\}$. The observations $\mathbf{x} = \{x_i, i = 0, ..., k\}$ are observed from data group. Bartholomew (1959) developed LRT for normal distributed data with variance σ^2. The LRT for binomial data is (Agresti and Coull, 1996)

$$LRT = \frac{SUP(L(\hat{g}(y|\mathbf{x}), y \in H_A^j))}{SUP(L(\hat{g}(y|\mathbf{x}), y \in H_0))} \approx \sum_{i=0}^{k} \frac{n_i(\widetilde{p}_i - \overline{p})^2}{\overline{p}(1-\overline{p})} \tag{3.16}$$

here, L is the likelihood function. $\overline{p} = \sum_0^k x_i / \sum_0^k n_i$ is the global mean estimator and $\widetilde{p}_i$ is the MLE for p_i under order restriction. $\widetilde{p}_i$s can be obtained by pool-adjacent-violators algorithm (PAVA) under Simple-order restriction:

$$\widetilde{p}_i = \min_{l \geq i} \max_{m \leq i} \frac{\sum_l^{j=m} x_j}{\sum_l^{j=m} n_j} \tag{3.17}$$

Since $\widetilde{p}_i$s are MLE, they can also be represented by $P_N(\mathbf{x}|H_A)$, which is least squares projection of the observation into the global alternative space H_A (Wright, 1988). From the definition of least squares projection, we know that vector $P_N(\mathbf{x}|H_A)$ has the smallest Euclidean norm from observation $\mathbf{x}$ to model H_A.

According to Robertson et al. (1988), the distribution is distributed as weighted chi-square, under H_0.

3.4.2 Single Contrast Test (SCT)

Before we introduce MCT, a binomial statistics of Single Contrast Test (SCT) is built for further description. A score function of the binomial statistics is defined as a standardized linear combination of the group sample means $\overline{p}_i$ divided by the pooled sample deviation estimator.

$$T = \frac{\sum_{i=0}^{k} \frac{c_i}{n_i} X_i}{\sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n_i}\}}} \tag{3.18}$$

Here, c_is are the contrast coefficients. $\sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n_i}\}}$ is the estimator of the variance under H_0. The test statistics T is asymptotic normal distributed, under H_0.

From the event $\{T \leq z_\alpha | H_0\}$ that T is smaller that the critical value under H_0, we can formulate a confidence interval, in which the value of the estimated proportion $\sum_{i=0}^{k} \frac{c_i}{n_i} X_i$ is likely to be included under H_0 with Type I error α.

$$\left(\sum_{i=0}^{k} \frac{c_i}{n_i} X_i - z_\alpha \sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n_i}}, \ \ +\infty \right] \tag{3.19}$$

This interval is one-sided, since we are focused on problem under order restriction.

Event that T is smaller that the critical value with Type I error α, is equivalent to event that zero lies in the confidence interval of the estimated proportion with Type I error α.

The Contrast Coefficients

The contrast coefficients c_is are constrained by $\sum_{i=0}^{k} c_i = 0$. This constrain will make the mean of all the test statistics T equal to zero if the the null hypothesis is true. We will show the proof of this shortly

$$
\begin{aligned}
E(T) &= E(\frac{\sum_{i=0}^{k} \frac{c_i}{n_i} X_i}{\sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n_i}\}}}) \\
&= E(\sum_{i=0}^{k} c_i X_i) * \frac{\sum_{i=0}^{k} \frac{1}{n_i}}{\sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n_i}\}}} \\
&= \{\sum_{i=0}^{k} c_i E(X_i)\} * \frac{\sum_{i=0}^{k} \frac{1}{n_i}}{\sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n_i}\}}}
\end{aligned} \tag{3.20}
$$

under the null hypothesis, all the X_i are from same distribution and they all have same expectation $E(X_i) = E(X)$

$$
E(T) = E(X) * \overbrace{\sum_{i=0}^{k} c_i}^{0} * \frac{\sum_{i=0}^{k} \frac{1}{n_i}}{\sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n_i}\}}} = 0 \tag{3.21}
$$

Besides the above constrain, c_is are free to be chosen.

Geometrically, the vector of contrasts $C = \{c_1, ..., c_k\}$ is orthogonal to the unit vector $\mathbf{1} = \{1, ..., 1\}$, because the linear product of these two vectors is zero $C \times \mathbf{1} = 0$. This means the contrasts vector will extract extra information, which describes how the observation is deviate from the center. Under balanced sample size situation, i.e. $n_i = n$, $i = 0, 1, ..., k$, equation 3.18 can be rewritten as:

$$
\begin{aligned}
T &= \frac{\sum_{i=0}^{k} \frac{c_i}{n} X_i}{\sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n}\}}} \\
&= \frac{\sum_{i=0}^{k} \frac{c_i}{n} X_i - \overline{p} n \overbrace{\sum_{i=0}^{k} \frac{c_i}{n}}^{0}}{\sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n}\}}} \\
&= \frac{\sum_{i=0}^{k} \frac{c_i}{n} (X_i - \overline{p} n)}{\sqrt{\{\overline{p}(1-\overline{p}) \sum_{i=0}^{k} \frac{c_i^2}{n}\}}}
\end{aligned} \tag{3.22}
$$

T is maximized if we choose an adaptive contrast coefficients with $c_i = X_i/n - \overline{p}$. Thus, c_i can be interpreted as the predictor of the differences between proportions X_i/n and $\overline{p}$. Generally, this is also true for unbalanced sample size situation (Bretz et al., 2005).

The Estimated Variance

This estimated variance $\sqrt{\{\overline{p}(1-\overline{p})\sum_{i=0}^{k}\frac{c_i^2}{n_i}\}}$ has poor behavior, e.g. when the proportion $\overline{p}$ is closed to 0, this variance will be very small. Then we lost the accuracy of T.

The idea of adjusted this problem comes from the justification of confidence interval. There are several alternative variance estimators, such as Add-4-method and Add-2-method, which can be used in adjusted the variance in many to one comparison (Schaarschmidt et al., 2009). Since reject the hypothesis by critical value and reject the hypothesis by correspond confidence interval are equivalent, we can use the adjust variance from confidence interval to calculate the value of T.

Here we just give a brief description of the Add-4-method. The Add-4-method for binomial proportions is invented by Agresti and Caffo (2000). The new proportion is achieved by adjusting the 2 by 2 table with two "successes and two "failures". i.e. $x_i' = x_i + 1$, $n_i' = n_i + 2$ and $\overline{p'} = \sum_0^k x_i' / \sum_0^k n_i'$.
The variance is adjusted to $\sqrt{\{\overline{p'}(1-\overline{p'})\sum_{i=0}^{k}\frac{c_i^2}{n_i'}\}}$.

3.4.3 Multiple Contrast Test (MCT)

Even mentioned by many older articles (for example Dunnett (1955) and Abelson and Tukey (1963)), the Multiple Contrast Test (MCT) was introduced by Mukerjee et al. (1987). Bretz and Hothorn (2002) also introduced the MCT method for binomial data. Here we will quote the result from Bretz and Hothorn (2002, 2003).

In SCT, for every vector of contrast coefficients $C_j = (c_{j,0}, ..., c_{j,k})$, there exists a correspond predict model H_A^j. By putting finite many prediction models together to

test the null hypothesis, we get the MCT, which can be defined as the maximum over certain chosen SCT. The test statistic is $T_{max} = MAX(T_1, T_2, ..., T_q)$. Under the null hypothesis, the test statistic $(T_1, T_2, ..., T_q)$ is q-variate central normal distributed under the null hypothesis and correlation the matrix is $R = \{\rho_{j,l}\}$ (detail will be given in Chapter 7).

The estimated correlation is:

$$\rho_{j,l} = \frac{\sum_{i=0}^{k} c_{j,i}c_{l,i}\hat{p}_i(1-\hat{p}_i)/n_i}{\sqrt{(\sum_{i=0}^{k} c_{j,i}^2\hat{p}_i(1-\hat{p}_i)/n_i)(\sum_{i=0}^{k} c_{l,i}^2\hat{p}_i(1-\hat{p}_i)/n_i)}} \tag{3.23}$$

The asymptotic power of multiple contrast tests is given by

$$\begin{aligned} & P(\max_{1\leqslant l\leqslant q}\{T_l\} \geqslant z_{q,1-\alpha} \mid H_A) \\ =& 1 - P(T_1 \leq z_{q,1-\alpha}, \text{and...and, } T_q \leq z_{q,1-\alpha} \mid H_A) \\ =& 1 - \Phi_q((z_{q,1-\alpha})\text{diag}(\frac{1}{\sqrt{V(T_1)}}, ..., \frac{1}{\sqrt{V(T_q)}}); \mathbf{e}, \mathbf{R}) \end{aligned} \tag{3.24}$$

Here, $z_{q,1-\alpha}$ is the q-variate normal $100(1-\alpha)$-equipercentage point under H_0. $e = \{(E(T_1), ..., E(T_q))\}$ and $v = \{(\sqrt{V(T_1)}, ..., \sqrt{V(T_q)})\}$ are the mean vector and the variance vector of $\{T_1, T_2, ..., T_q\}$ under H_A (Bretz and Hothorn, 2002).

3.4.4 Cochran-Armitage Test (CAT)

The Cochran-Armitage Test is a linear weighted regression test, which is locally most powerful test if some extra information is known. Here we use the results from Agresti (2002). A linear model is assumed as

$$p_i = \alpha + \beta w_i \tag{3.25}$$

which will be fitted by weighted least squares. Here, $\mathbf{w} = \{w_i, i = 0, ..., k\}$ is the score vector, which describes the distances between the treatments, e.g. dose level. The null hypothesis of this independence test is $H_0 : \beta = 0$. $\bar{w} = \sum_0^k n_i w_i / \sum_0^k n_i$ is

the average score. We can apply linear regression and predict the true value p_i as

$$\hat{p}_i = \overline{p} + \hat{\beta}(w_i - \bar{w}) \tag{3.26}$$

where

$$\hat{\beta} = \frac{\sum_0^k n_i(\overline{p}_i - \overline{p})(w_i - \bar{w})}{\sum_0^k n_i(w_i - \bar{w})^2} \tag{3.27}$$

The test statistics for Cochran-Armitage Test is

$$CAT = \frac{\hat{\beta}^2 \sum_{i=0}^k n_i(w_i - \bar{w})^2}{\overline{p}(1-\overline{p})} \tag{3.28}$$

which is chi-squared distributed with one degree of freedom. The critical values are $z_{1-\alpha} = \chi^2_{1,1-\alpha}$ for two-sided test.

For our order restriction problem, we can also restrict the slope parameter β to be strict positive, i.e. $\beta' = max(\beta, 0)$. When $\beta > 0$

$$CAT' = \frac{\hat{\beta'}^2 \sum_{i=0}^k n_i(w_i - \bar{w})^2}{\overline{p}(1-\overline{p})} \tag{3.29}$$

When $\beta \leq 0$, $CAT' = 0$

The critical value can be calculated as $z_{1-\alpha} = \chi^2_{1,1-2\alpha}$

3.4.5 Advantages and disadvantages of test methods

The advantage of these test methods is that LRT has a good "average" power over the global alternative. If the contrasts are "strategically" chosen within the global alternative space, MCT can achieve a good power too. (Bretz, 1999)

However, MCT and LRT are not designed to do model selection. The PAVA estimator which maximizes LRT is the MLE for the global alternative. LRT only tests

the global alternative. It is impossible to select the model from local alternatives. The estimator which maximize the MCT is the MLE for the local alternative. But MCT selects the model which has the MLE or in other words "Best fits the data", as the best model. As we discussed in the previous section, model selection should achieve the balance between the model fitting and complexity. Bretz et al. (2005) described the over fitting problem to use MCT for model selection and suggested Information Criterion method for model selection.

The disadvantage for Cochran-Armitage Test (CAT) is obvious that it is a trend test and it need extra information to be the most powerful test. Furthermore, CAT cannot do model selection after the test. In Chapter 8, we will show that it is not a suitable method to solve our problems.

Chapter 4

Information Criterion (IC) for model selection

In Chapter 3, we have already discussed about what a good model is. In this chapter, we will introduce Kullback-Leibler (KL) distance which measures the differences between the estimated model and the true model. However, the calculation of KL distance is not easy. Usually people use adjusted maximum likelihood to approximate it (Akaike, 1974). In the following sections, we will also give an algorithm to minimize the distance and calculate the adjustment of the distance under different situations. Finally, we will discuss about Akaike Information Criterion (AIC) for general situations, Order Restricted Information Criterion (ORIC) for one-sided order restriction and a new Information Criterion, which considers the complexity during distance calculation and adjusts the bias. In order to help the readers understand the procedure better, we also give a short and brief calculation of the problems mentioned in the former chapter.

4.1 Kullback-Leibler (KL) distance

Kullback-Leibler (KL) distance is a measurement for the Information distance between the estimated model density function $\hat{g}(y)$ and the true model density function

$g(y)$(Anraku, 1999).

$$KL(g(y), \hat{g}(y)) = \int g(y) * \log g(y) dv(y) - \int g(y) * \log \hat{g}(y) dy \geq 0 \tag{4.1}$$

The equality is achieved if and only if $g(y) = \hat{g}(y)$ almost surely. The first term is a constant which is always lager than or equal to the second term. Then we have

$$KL(g(y), \hat{g}(y|\hat{\theta}(x))) = \text{Constant} - \int g(y) * \log \hat{g}(y|\hat{\theta}(x)) dy \tag{4.2}$$

here, $\hat{g}(y|\hat{\theta}(x))$ is the estimated distribution function given by observed data $x = \{x_1, x_2, ...x_n\} \in X$, X is the set of all the observations and $\hat{\theta} = \{\hat{\theta}_1, \hat{\theta}_2, ...\hat{\theta}_m, m < n\}$ is the estimated parameter of the true parameter θ. The second term includes the true distribution which is unknown. It cannot be calculated, but can be estimated by the maximum likelihood estimator $\tilde{\theta}$

$$KL(g(y), \hat{g}(y|\tilde{\theta}(x)) \approx \text{Constant} - \log \hat{g}(x|\tilde{\theta}(x)) \tag{4.3}$$

4.1.1 Bias of the Log-Likelihood

In this section, we will calculate the bias between the estimated KL distance and the true KL distance. The readers should notice that $\tilde{\theta} \xrightarrow{n} \theta$ is not a sufficient condition for

$$f(\tilde{\theta}) \xrightarrow{n} f(\theta) \tag{4.4}$$

Even if we have unbiased estimators of all the parameters, the KL distance, which is calculated by these estimators, are still biased. Many people have discussed about the method to calculate the bias under different situations (Akaike, 1974; Akaike and Kitagawa, 1998; Anraku, 1999; Burnham and Anderson, 2004; Hughes and King, 2003; Konishi and Kitagawa, 2008). We put the MLE $\tilde{\theta}$ in Equation 4.3 and Equation 4.2. Then we use Equation 4.3 to substitute Equation 4.2. Now we have the bias, which is caused by $\tilde{\theta}$, as the expected value of the difference (Konishi and Kitagawa,

2008; Hughes and King, 2003)

$$\begin{aligned} & Bias(g(y), \hat{g}(y|\tilde{\theta}(x))) \\ =& E_X\{\text{Constant} - \int g(y) * \log \hat{g}(y|\tilde{\theta}(x))dy\} - E_X\{\text{Constant} - \log \hat{g}(x|\tilde{\theta}(x))\} \\ =& E_X\{\log \hat{g}(x|\tilde{\theta}(x)) - \int g(y) * \log \hat{g}(y|\tilde{\theta}(x))dy\} \end{aligned} \tag{4.5}$$

This difference can be separated into 3 parts,

$$\begin{aligned} Bias(g(y), \hat{g}(y|\tilde{\theta}(x))) =& E_X\{\log \hat{g}(x|\tilde{\theta}(x)) - \log g(x|\theta(x))dy\} \\ &+ E_X\{\log g(x|\theta(x)) - \int g(y) * \log g(y|\theta(x))dy\} \\ &+ E_X\{\int g(y) * \log g(y|\theta(x))dy - \int g(y) * \log \hat{g}(y|\tilde{\theta}(x))dy\} \\ =& D_1 + D_2 + D_3 \end{aligned} \tag{4.6}$$

Calculation of $\mathbf{D}_3$

Let the second term of Equation 4.2 be a function of parameter ϑ', we have

$$\eta(\theta') = \int g(y) * \log g(y|\theta'(x))dy \tag{4.7}$$

Let $\theta' = \tilde{\theta}$ we make a Taylor expansion of $\eta(\tilde{\theta})$ around the true parameter θ

$$\begin{aligned} \eta(\tilde{\theta}) =& \eta(\theta) + \Sigma_{i=1}^{m}(\tilde{\theta}_i - \theta_i)\frac{\partial \eta(\theta)}{\partial_i} \\ &+ \frac{1}{2}\Sigma_{i=1}^{m}\Sigma_{j=1}^{m}(\tilde{\theta}_i - \theta_i)(\tilde{\theta}_j - \theta_j)\frac{\partial^2 \eta(\theta)}{\partial_i \partial_j} + ... \end{aligned} \tag{4.8}$$

here $\tilde{\theta} = \{\tilde{\theta}_1, \tilde{\theta}_2, ...\tilde{\theta}_m\}$ and $\theta = \{\theta_1, \theta_2, ...\theta_m\}$. Because θ is the true parameter, so it maximizes function $\eta(\tilde{\theta})$, i.e. $\frac{\partial \eta(\theta)}{\partial_i} = 0$. Now we have

$$\begin{aligned} \eta(\tilde{\theta}) \approx& \eta(\theta) + \frac{1}{2}\Sigma_{i=1}^{m}\Sigma_{j=1}^{m}(\tilde{\theta}_i - \theta_i)(\tilde{\theta}_j - \theta_j)\frac{\partial^2 \eta(\theta)}{\partial_i \partial_j} \\ =& \eta(\theta) + \frac{1}{2}(\tilde{\theta} - \theta)^t H(\theta')|_{\theta'=\theta}(\tilde{\theta} - \theta) \end{aligned} \tag{4.9}$$

$H(\theta')$ is the Hessian matrix of $\eta(\theta')$. Put this approximation of $\eta(\tilde{\theta})$ back in to D_3, we have

$$\begin{aligned} D_3 =& E_X \int g(y) * \log g(y|\theta(x)) dy - \int g(y) * \log \hat{g}(y|\tilde{\theta}(x)) dy\} \\ =& E_X \{\eta(\theta) - \eta(\theta) - \frac{1}{2}(\tilde{\theta} - \theta)^t H(\theta')|_{\theta'=\theta}(\tilde{\theta} - \theta)\} \\ =& \frac{1}{2}(\tilde{\theta} - \theta)^t I(\theta')|_{\theta'=\theta}(\tilde{\theta} - \theta) \end{aligned} \tag{4.10}$$

where I is the information matrix.

Calculation of $\mathbf{D}_2$

The calculation of D_2 is quite simple, since no estimation here

$$\begin{aligned} D_2 =& E_X \{\log g(x|\theta(x)) - \int g(y) * \log g(y|\theta(x)) dy\} \\ =& E_X \{\log g(x|\theta(x)) - E_Y \{\log g(x|\theta(x)) \\ =& 0 \end{aligned} \tag{4.11}$$

In some references (e.g. Hughes and King (2003)), D_2 and D_3 are considered together as the bias caused by the MLE $\tilde{\theta}$, while D_1 is considered as the bias caused by the estimated function.

Calculation of $\mathbf{D}_1$

Similar as what we have done to calculate D_3, Taylor expansion will be used to get the approximation. Let $l(\theta') = \log g(x|\theta')$, we expand $l(\theta')$ around the MLE $\tilde{\theta}$

$$\begin{aligned} l(\theta') =& l(\tilde{\theta}) + \Sigma_{i=1}^m (\theta'_i - \tilde{\theta}_i) \frac{\partial l(\tilde{\theta})}{\partial_i} \\ & + \frac{1}{2} \Sigma_{i=1}^m \Sigma_{j=1}^m (\theta'_i - \tilde{\theta}_i)(\theta'_j - \tilde{\theta}_j) \frac{\partial^2 l(\tilde{\theta})}{\partial_i \partial_j} + ... \end{aligned} \tag{4.12}$$

Since $\tilde{\theta}$ is the MLE we have $\frac{\partial l(\tilde{\theta})}{\partial \theta'} = 0$. Similar as what we have calculated D_3, the value of $l(\theta)$ can be estimate by $l(\tilde{\theta})$ as

$$\begin{aligned} l(\theta) \approx & l(\tilde{\theta}) + \frac{1}{2}\Sigma_{i=1}^{m}\Sigma_{j=1}^{m}(\theta_i - \tilde{\theta}_i)(\theta_j - \tilde{\theta}_j)\frac{\partial^2 l(\tilde{\theta})}{\partial_i \partial_j} \\ = & l(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})^t H(\theta')|_{\theta'=\tilde{\theta}}(\theta - \tilde{\theta}) \end{aligned} \tag{4.13}$$

here $H(\theta')$ is the Hessian matrix of $l(\theta')$. Put this approximation back in to D_1, we have

$$\begin{aligned} D_1 = & E_X\{\log \hat{g}(x|\tilde{\theta}(x)) - \log g(x|\theta(x))dy\} \\ = & \frac{1}{2}(\tilde{\theta} - \theta)^t I(\theta')|_{\theta'=\tilde{\theta}}(\tilde{\theta} - \theta) \end{aligned} \tag{4.14}$$

where I is the information matrix.

Finally we have the bias as

$$\begin{aligned} & D_1 + D_2 + D_3 \\ = & E_X\{\log \hat{g}(x|\tilde{\theta}(x)) - \log g(x|\theta(x))dy\} \\ & + E_X\{\log g(x|\theta(x)) - \int g(y) * \log g(y|\theta(x))dy\} \\ & + E_X\{\int g(y) * \log g(y|\theta(x))dy\} - \int g(y) * \log \hat{g}(y|\tilde{\theta}(x))dy\} \\ = & \frac{1}{2}(\tilde{\theta} - \theta)^t I(\theta')|_{\theta'=\tilde{\theta}}(\tilde{\theta} - \theta) + 0 + \frac{1}{2}(\tilde{\theta} - \theta)^t I(\theta')|_{\theta'=\theta}(\tilde{\theta} - \theta) \\ = & \frac{1}{2}(\tilde{\theta} - \theta)^t I(\tilde{\theta})(\tilde{\theta} - \theta) + \frac{1}{2}(\tilde{\theta} - \theta)^t I(\theta)(\tilde{\theta} - \theta) \\ \approx & (\tilde{\theta} - \theta)^t I(\tilde{\theta})(\tilde{\theta} - \theta) \end{aligned} \tag{4.15}$$

These two terms are asymptotic equivalent. In later part of this thesis, we consider them as the same information term. Under different conditions, the information term is different. In next sections, we will introduce One-sided AIC (OSAIC), which is developed by Hughes and King (2003). AIC and ORIC can also be generated as a special case of KL distance too.

4.1.2 One-sided AIC (OSAIC) and ORIC

In this section, we will introduce One-sided AIC (OSAIC), which is developed by Hughes and King (2003). The twice of the information term mentioned in last section, i.e.

$$2(\tilde{\theta}-\theta)^t I(\theta')|_{\theta'=\tilde{\theta}}(\tilde{\theta}-\theta) \tag{4.16}$$

is "asymptotically equivalent to the distribution under H_0 the partially inequality constrained Wald test statistics" (Hughes and King, 2003). Under elementary alternative model

$$H_A^j : p_1 = p_2 = ... < p_{j_1} = ... < p_{j_r} = ... = p_k,\ j = 1, ..., \frac{k!}{r!(k-r)!},\ 1 < j_1 < ... < j_r < k \tag{4.17}$$

with exact r inequality constrains, twice of the information term is asymptotic weighted chi-square distributed with $k - r + m$ degrees of freedom $\sum_{m=0}^{r} w(r,m)\chi^2(k - r + m)$, where $w(r,m)$ are weighted probability i.e. $\sum_{m=0}^{r} w(r,m) = 1$. We can use the level probability developed by Robertson et al. (1988) to calculate this weighted probability as following

$$w(r,m) = P\{r, m, \omega\{H_A^j\}\} \tag{4.18}$$

For any value of r, we can calculate the bias as

$$\begin{aligned} Bias(g(y), \hat{g}(y|\tilde{\theta}(x))) =& D_1 + D_2 + D_3 \\ =& \frac{1}{2}(\tilde{\theta}-\theta)^t I(\theta')|_{\theta'=\tilde{\theta}}(\tilde{\theta}-\theta) + \frac{1}{2}(\tilde{\theta}-\theta)^t I(\theta')|_{\theta'=\theta}(\tilde{\theta}-\theta) \\ \approx& (\tilde{\theta}-\theta)^t I(\theta')|_{\theta'=\tilde{\theta}}(\tilde{\theta}-\theta) \\ =& E(\sum_{m=0}^{r} w(r,m)\chi^2(k-r+m)) \\ =& \sum_{m=0}^{r} w(r,m)(k-r+m) \end{aligned} \tag{4.19}$$

We have the OSAIC under alternative model H_A^j as

$$OSAIC(H_A^j) = \log \hat{g}(x|\tilde{\theta}(x)) - \sum_{m=0}^{r} w(r,m)(k-r+m) \tag{4.20}$$

For binomial data we have

$$OSAIC(H_A^j) = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i-x_i)!}(\hat{p}_i^{x_i} * (1-\hat{p}_i)^{n_i-x_i}) - \sum_{m=0}^{r} w(r,m)(k-r+m) \tag{4.21}$$

and another common form is achieved by multiplying "-2" to the former one

$$OSAIC(H_A^j) = -2logL(model) + 2\sum_{m=0}^{r} w(r,m)(k-r+m) \tag{4.22}$$

When $r = k$, the last model

$$H_A^{2^{k-1}} : p_1 < p_2 < ... < p_k, \tag{4.23}$$

has exactly k inequality. The bias is reduced to

$$\begin{aligned} Bias(g(y), \hat{g}(y|\tilde{\theta}(x))) &= \sum_{m=0}^{r} w(r,m)(k-r+m) \\ &= \sum_{m=0}^{k} w(k,m)m \end{aligned} \tag{4.24}$$

which is the bias term developed by Anraku (1999). For binomial data we have

$$ORIC = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i-x_i)!}(\hat{p}_i^{x_i} * (1-\hat{p}_i)^{n_i-x_i}) - \sum_{m=0}^{k} w(k,m)m \tag{4.25}$$

When $r = 0$ the bias is reduced to

$$\begin{aligned} Bias(g(y), \hat{g}(y|\tilde{\theta}(x))) &= \sum_{m=0}^{r} w(r,m)(k-r+m) \\ &= k \end{aligned} \tag{4.26}$$

which is the bias term developed by Akaike (1974).

4.1.3 Partition sets and its estimators

In the last section, we have used level probability to calculate the ORIC. In this section, we will introduce partition sets and will show how to calculate level probability by using partition sets.

Definition

The partition sets estimator $\check{\theta}(x)$ is the pooled mean within each partition sets, which has already been discussed in previous chapters. It is a very important statistic to calculate local MLE and SLE.

Under different models, the regions, which are separated by inequalities, are called Partition sets. The partition sets estimator $\check{\theta}$ is the pooled mean within each partition set. For example, under Single Change-point model, alternative model H_A^j has two partition sets, $p_0 = ... = p_{j-1}$ and $p_j = ... = p_k$. The partition sets estimators are

$$\begin{aligned} \check{p}_0 = ... = \check{p}_{j-1} &= \frac{\sum_{i=0}^{j-1} x_i}{\sum_{i=0}^{j-1} n_i} \\ \check{p}_j = ... = \check{p}_k &= \frac{\sum_{i=j}^{k} x_i}{\sum_{i=j}^{k} n_i} \end{aligned} \tag{4.27}$$

Under Simple-order restriction, the last alternative model $H_A^{2^k-1} : p_0 < ... < p_{j-1} < p_j < ... < p_k$, has k partition sets and each partition set only has one parameter. The partition sets estimators are just the same as the simple estimator

$$\check{p}_j = \frac{x_j}{n_j} \tag{4.28}$$

Hypothesis	No. of Partition	$\check{p}_0$	$\check{p}_1$	$\check{p}_2$
H_0: $p_0 = p_1 = p_2$	1	0.5000	0.5000	0.5000
H_A^1: $p_0 < p_1 = p_2$	2	0.4500	0.5119	0.5119
H_A^2: $p_0 = p_1 < p_2$	2	0.4444	0.4444	0.5854
H_A^3: $p_0 < p_1 < p_2$	3	0.4500	0.4400	0.5854

Table 4.1: The partition sets under Simple-order restriction

Example

Let us combine the two kinds of models we mentioned above to have a mixed alternative model and give a real example. In our previous example of adverse events, we want to test if there is a Simple-order restriction. For $k = 2$, the alternative model are mixed by three elementary models, $H_A = H_A^1 \bigcup H_A^2 \bigcup H_A^{2^k-1}$. Here H_A^1 and H_A^2 are elementary alternative models for Single Change-point. $H_A^{2^k-1} = H_A^3$ is elementary alternative models for Simple-order. The partition sets under Simple-order restriction are given in Table 4.1

4.1.4 Level probability

The level probability $P\{k, l, \omega\{H_A^j\}\}$ is developed by Robertson et al. (1988) under Simple-order and Simple-tree order restriction. Hughes and King (2003) extended it to weighted probability $w(r, m)$ for all kinds of one-sided order restriction. Here we just give a short and brief description of the numerical calculation algorithm developed by Robertson et al. (1988) for normal data and by Hothorn et al. (2008) for binomial data.

The level probability $P\{k, l, \omega\{H_A^j\}\}$ under model H_A^j can be defined as following: Given k random variables $\{Y_1, Y_2, ..., Y_k\} \in H_A^j$, $P\{k, l, \omega\{H_A^j\}\}$ is the probability that theses variables can be divided into l partition sets.

For example, for binomial data, let $k = 2$, given $\{Y_1 = \frac{x_1}{n_1}, Y_2 = \frac{x_2}{n_2}\} \in H_A^{2^k-1}$, where $H_A^{2^k-1} = \{Y_1 < Y_2\}$ is under the Simple-order restriction for dimension two, the level

probabilities are

$$P\{2,2,\omega\{H_A^j\}\} = P\{Y_1 < Y_2\} = P\{1,2,\omega\{H_A^j\}\} = P\{Y_1 \geq Y_2\} = \frac{1}{2} \tag{4.29}$$

For binomial data, let $k = 3$, given $\{Y_1 = \frac{x_1}{n_1}, Y_2 = \frac{x_2}{n_2}, Y_3 = \frac{x_3}{n_3}\} \in H_A^{2^k-1}$, where $H_A^{2^k-1} = \{P_1 < P_2 < ... < P_k\}$ is the total Simple-order restriction, the level probabilities are (Robertson et al., 1988)

$$\begin{aligned} & P\{3,3,\omega\{H_A^j\}\} \\ =& P\{Y_1 < Y_2 < Y_3\} \\ =& \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho) \end{aligned} \tag{4.30}$$

here $\rho = \sqrt{\frac{n_1 n_3}{(n_1+n_2)(n_2+n_3)}}$.

$$\begin{aligned} & P\{2,3,\omega\{H_A^j\}\} \\ =& P\{Y_1 = Y_2 < Y_3\} + P\{Y_1 < Y_2 = Y_3\} \\ =& P\{\frac{x_1+x_2}{n_1+n_2} < Y_3\}P\{2,2,\omega\{H_A^j\}\} + P\{Y_1 < \frac{x_2+x_3}{n_2+n_3}\}P\{2,2,\omega\{H_A^j\}\} \\ =& \frac{1}{2} \\ & P\{1,3,\omega\{H_A^j\}\} \\ =& 1 - P\{2,3,\omega\{H_A^j\}\} - P\{3,3,\omega\{H_A^j\}\} \\ =& \frac{1}{4} - \frac{1}{2\pi}\arcsin(\rho) \end{aligned} \tag{4.31}$$

4.1.5 Ninomiya AIC (NIC)

Note the number of Change-points as m. The AIC and ORIC only consider the bias term caused by unknown parameters, p_i. If the sequence is long enough($k/m > 3$), the bias term caused by Change-points should also be taken into account. Ninomiya

(2005) considered the position of Change-point as extra unknown parameter and used the information of this to make a better approximation of the Taylor Expansion. The calculation of the information matrix is quite complicate. The readers can refer to his paper for detail.

Under epidemic Change-points order restriction, the NIC for different hypotheses are

$$\begin{aligned} NIC(H_0) =& log(L(\hat{g}(x|\tilde{\theta}(x)))) - 1 \\ =& -\{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!}(\tilde{p}_{j,i}^{x_i} * (1 - \tilde{p}_{j,i})^{n_i - x_i})\} - 1 \end{aligned} \tag{4.32}$$

$$\begin{aligned} NIC(H_A^j) =& log(L(\hat{g}(x|\tilde{\theta}_j(x)))) - 2 - 3 * m \\ =& -\{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!}(\tilde{p}_{j,i}^{x_i} * (1 - \tilde{p}_{j,i})^{n_i - x_i})\} - 2 - \text{penalty} \end{aligned} \tag{4.33}$$

According to Ninomiya (2006), we use the following equation to calculate the penalty term for binomial data.

Bias for one Change-point from $B(n, p^{(1)})$ to $B(n, p^{(2)})$ is

$$b(p^{(1)}, p^{(2)}) = \frac{c_1^2\sigma_2^4 + c_1c_2\sigma_1^2\sigma_2^2 + c_2^2\sigma_1^4}{c_1c_2(c_1\sigma_2^2 + c_2\sigma_1^2)}$$

where

$$c_1 = n \log \frac{1 - p^{(1)}}{1 - p^{(2)}} - np^{(1)} \log \frac{p^{(2)}(1 - p^{(1)})}{p^{(1)}(1 - p^{(2)})}$$

$$c_2 = n \log \frac{1 - p^{(2)}}{1 - p^{(1)}} - np^{(2)} \log \frac{p^{(1)}(1 - p^{(2)})}{p^{(2)}(1 - p^{(1)})}$$

$$\sigma_1^2 = np^{(1)}(1-p^{(1)})\log\frac{p^{(2)}(1-p^{(1)})}{p^{(1)}(1-p^{(2)})}^2$$

$$\sigma_2^2 = np^{(2)}(1-p^{(2)})\log\frac{p^{(2)}(1-p^{(1)})}{p^{(1)}(1-p^{(2)})}^2$$

Therefore, NIC for Epidemic-order ($\mathrm{p}^{(1)} = ... = p^{(1)} > p^{(2)} = ... = p^{(2)} < p^{(1)} = ... = p^{(1)}) is$

$$\log L(x|\widetilde{p}^{(1)}, \widetilde{p}^{(2)}) - 2 - b(\widetilde{p}^{(1)}, \widetilde{p}^{(2)}) - b(\widetilde{p}^{(2)}, \widetilde{p}^{(1)})$$

the term "2" is because there are two unknown parameter.

4.1.6 Akaike Information Criterion (AIC)

In this section, we generate AIC from KL distance by simplifying the penalty term. We want readers to notice that AIC is a simple approximation of KL distance. The index i is from 0 to k. So here we have one more unknown parameter than before. The penalty term is changed from k to $k+1$.

Akaike (1974) used Taylor expansion to estimate the second term as log-likelihood minus bias adjustment term, which is equal to the number of unknown parameters. The adjustment term is always positive, so it is also called penalty term.

$$\begin{aligned} AIC =& -E_x(\int g(y) * \log\hat{g}(y|\hat{\theta}(x))dy) \\ =& -\{\log\hat{g}(x|\tilde{\theta}(x)) + \text{TaylorExpansion}(\hat{g}(x|\tilde{\theta}(x))\} \\ \approx& -\{\log\hat{g}(x|\tilde{\theta}(x)) - \mathbf{Penalty}(\hat{g}(x|\tilde{\theta}(x))\} \\ =& -\log\hat{g}(x|\tilde{\theta}(x)) + \{\mathbf{Penalty} = \mathbf{m}\} \end{aligned} \tag{4.34}$$

Here estimator $\hat{\theta}(x)$ is the MLE $\tilde{\theta}(x)$. If true density function $g(y)$ has normal distribution $N(u_i, \sigma), i = 0, ..., k$ and $\hat{g}(y)$ is the estimated density function which

has normal distribution $N(\hat{u}_i, \sigma), i = 0, ..., k$, the expected distance between them is (Konishi and Kitagawa, 2008)

$$\begin{aligned} & E_x(KL(g(y), \hat{g}(y))) \\ =&\text{Constant} - E_x(\int g(y) * \log \hat{g}(y) dy) \\ \approx&\text{Constant} - \{\log \prod_{\mathrm{i=0}}^{\mathrm{k}} \frac{1}{2\pi\sigma^2} \exp\{\frac{-(x_i - \hat{u}_i)^2}{2\sigma^2}\} - \text{penalty}(\hat{g})\} \\ =&\text{Constant}' - \{-\frac{1}{2}\sum_{i=0}^{k} \frac{(x_i - \hat{u}_i)^2}{\sigma^2} - (k+1)\} \end{aligned} \tag{4.35}$$

If true density function $g(y)$ has binomial distribution $B(p_i, n_i), i = 0, ..., k$ and $\hat{g}(y)$ is the estimated density function which has binomial distribution $B(\hat{p}_i, \hat{n}_i), i = 0, ..., k$, then we have the expected KL distance as

$$\begin{aligned} & E_x(KL(g(y), \hat{g}(y))) \\ =&\text{Constant} - E_x(\int g(y) * \log \hat{g}(y) dy) \\ \approx&\text{Constant} - \{\log \prod_{i=0}^{k} \frac{n_i!}{x_i!(n_i - x_i)!} (\hat{p}_i^{x_i} * (1 - \hat{p}_i)^{n_i - x_i}) - \text{penalty}(\hat{g})\} \\ =&\text{Constant} - \{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\hat{p}_i^{x_i} * (1 - \hat{p}_i)^{n_i - x_i}) - (k+1)\} \end{aligned} \tag{4.36}$$

Based on entropy, AIC offers a measurement which balances "goodness of fit" and complexity for statistic modeling. It builds connections between likelihood and information criterion. Many people give improvement of it, such as NIC, ORIC and OSAIC. The ICs based on AIC build a big "family". Here we just have introduced some members of them.

The disadvantages of these methods, which use information criterion, are that theses methods do not control the FWER. They are designed for model selection other than testing.

4.2 Improve the maximum likelihood estimator by penalty term

Usually it is impossible to calculate the KL distance, because the "true" distribution is unknown. Therefore, different types of information criterion, such as AIC for general situation and ORIC for order restriction, are developed to estimate the expectation of the KL distance. They are adjusted by different to achieve the unbiased estimation of the KL distance.The model, which has the smallest KL distance to the true model, is selected as the most possible model.

In order to calculate the smallest KL distance, we need to maximize the difference between the log-likelihood and penalty term of different models. The maximum likelihood estimator (MLE) can be calculated under different situations, such as order restriction. The MLE is not hard to calculate by analytical method or numerical method, but the penalty terms usually depend on the MLE and its calculating algorithms. Furthermore, if we use too many parameters to fit the data, the penalty term will be so large that the effort achieved by increasing the number of parameters is totally "underperformed"(Zucchini, 2000).

AIC is the most well known one for model selection and is simple and easy to achieve. However, it has disadvantage that the extra Information from the structure and sample sizes is not used. The original Taylor expansion which is a function of the full model, is simplified by AIC only as the number of unknown parameters.

Furthermore, the MLE, which maximizes AIC, might not maximize the true KL distance. Because the KL distance depends on both of the maximum likelihood and the penalty term. Under certain order restriction, such as under Simple-order restriction, the penalty term for some alternative hypotheses could be very large, if too many parameters are estimated. Noticing that the penalty term is always positive, we realize that the MLE used by Akaike (1974); Anraku (1999) has too many degrees of freedom and becomes biased in such situation. Improved statistic models which concur such "underperformance" are introduced. In these models,

only one variable, which describes the total divergence between the partition sets and the overall mean, is considered. So the whole models have only two degrees of freedom.

4.2.1 Notification of the global, local and Suitable Likelihood Estimators

In the following sections of this chapter, we will use five kinds of estimators to calculate the likelihood. So here we list them together for readers to take a brief review of them. The most important one is the Partition sets estimator (PE), which is a very important statistics to calculate the Maximum likelihood estimator (MLE) and the Suitable likelihood estimator (SLE). We will also explain what is the difference between global MLE and local MLE. The global MLE, which can be calculated by PAVA, is treated as the projection of observed data in the alternative space H_A, while local MLEs can be treated as the projections of observed data in the alternative spaces H_A^j. Local MLE is more reasonable to use in model selection than the traditional global MLE.

Let the general estimated parameter (GEP) $\hat{\theta}(x)$ present the estimated parameter with some given statistic models and θ presents the true parameter. $\hat{\theta}(x)$ is estimated from the observed data x. It can be chosen as MLE, SLE or any other estimators. Let $\hat{g}(x|\hat{\theta}(x))$ present the estimated density function of the true distribution $g(x|\theta)$ with given estimator $\hat{\theta}(x)$.

Global MLE

The global Maximum likelihood estimator (gMLE), $\tilde{\theta}(x)$, is one of the most important kinds of estimators. It maximizes the likelihood of the estimation function $\hat{g}(x|\hat{\theta})$. So we have

$$\tilde{\theta} = \arg\max_{\hat{\theta}} \hat{g}(x|\hat{\theta}) \tag{4.37}$$

An algorithm, which calculates the gMLE $\tilde{\theta} = \{\tilde{p_0}, ..., \tilde{p_k}\}$ under Simple-order restriction, is given by Robertson et al. (1988) for normal distributed data.

As we have already discussed in last section, the log-likelihood for binomial data is

$$\log L(H) = \{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (p_i^{x_i} * (1 - p_i)^{n_i - x_i})\} \tag{4.38}$$

Hothorn et al. (2008) calculated the gMLE for binomial data as

$$\tilde{p_i} = \min_{l \geq i} \max_{m \leq i} \frac{\sum_l^{j=m} x_j}{\sum_l^{j=m} n_j} \tag{4.39}$$

The maximum log-likelihood is calculated by putting the gMLE into Equation 4.38

$$\log L(H) = \{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\tilde{p}_i^{x_i} * (1 - \tilde{p}_i)^{n_i - x_i})\} \tag{4.40}$$

Under Simple-order restriction, Anraku (1999) and Hughes and King (2003) used this global MLE for all elementary models to achieve the maximum likelihood. Simulation shows that this algorithm is "underperformed" when k is larger than two (Hothorn et al., 2008). Here we introduced our algorithm, which calculates MLE under each given elementary alternatives. This local MLE also reduces the "underperformce". After introducing all the model selection method, we will give a short and brief comparison to clear this. As we will see in Table 4.5, a simulation study is given to compare ORIC method with different likelihood estimators.

Local MLE

In this subsection, we will show how to calculate the local MLE (lMLE) and how is the relationship between lMLE and gMLE.

We can get the local MLE in the following steps. First, the partition sets estimator $\check{\theta}_j(x) = \{\check{p}_{j,0}, ..., \check{p}_{k,j}\}$ for each elementary alternative model H_A^j is calculated. Then

we order the pooled mean estimator of each partition sets. This estimator has full order restriction per definition. Then the local MLE is calculated by these partition sets estimator

$$\tilde{p}_{i,j} = \min_{l \geq i} \max_{m \leq i} \frac{\sum_{l}^{v=m} \check{p}_{j,v}}{\sum_{l}^{v=m} n_v} \tag{4.41}$$

put it in to the log-likelihood equation, we have

$$\log L(H_A^j) = \{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\tilde{p}_{i,j}^{x_i} * (1 - \tilde{p}_{i,j})^{n_i - x_i})\} \tag{4.42}$$

Without the loss of generality, all the MLE in later part of the this thesis is referred to local MLE. The ORIC for binomial data listed in this thesis is an improved version of the old one developed by Hothorn et al. (2008).

Suitable likelihood estimator

The Suitable likelihood estimator (SLE), $\bar{\theta}(x)$, is another important kind of estimator. It is designed to obtain the unbiased KL distance to concur the "underperformance". Even strong evidence has been shown in simulation study. We still category the prove of unbiasness as our third open question.

Let $\hat{p} = \sum_{i=0}^{k} (\frac{x_i}{n_i})$ to be the overall mean. The total difference is defined as

$$\Delta = \frac{\sum_{i=0}^{k} |x_i - \hat{p} n_i|}{1/(k+1) * \sum_{i=0}^{k} n_i} \tag{4.43}$$

For the null hypothesis, the "suitable" likelihood estimators (SLE) are the same as the MLE for AIC.

$$\bar{p} = \tilde{p} = \frac{\sum_{i=0}^{k} x_i}{\sum_{i=0}^{k} n_i} \tag{4.44}$$

The SLEs for given alternative hypotheses H_A^j are changed to

$$\bar{p}_{i,j} = \hat{p} + \frac{c_{j,i}}{C_j} * \Delta \tag{4.45}$$

here, $c_{i,j}$ is the contrast coefficients for MCT and $C_j = |\sum_{i=0}^{k} c_{j,i}|$. By doing this we have built a model selection method which shares the estimators with MCT. We are trying to build a bridge between model selection and test. Further discussion for this will be given later.

Put the SLE in to the log-likelihood equation, we have

$$\log L(H_A^j) = \{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\bar{p}_{i,j}^{x_i} * (1 - \bar{p}_{i,j})^{n_i - x_i})\} \tag{4.46}$$

The advantage to use SLE is that the degrees of freedom for log-likelihood under all elementary alternatives are equal to two. We can use this property to calculate the penalty terms, which are all the same.

Relationship of these estimators

Simply to prove per definition, the log-likelihood of these estimators has the following relationship.

$$\log L(gMLE) \geq \log L(lMLE) \geq \log L(SLE) \tag{4.47}$$

where the equality only achieves when the estimators are totally the same. This means the projection points are overlapped. However, this result does not show that gMLE is the best among those. In the simulation study section, we will see that

gMLE is sometimes the worst estimator, which is unstable and "overperformed".

4.2.2 Estimators under Single Change-point order restriction

Under Single Change-point order restriction, the MLE for the null model H_0 is

$$\tilde{p}_{0,0} = ... = \tilde{p}_{0,k} = \frac{\sum_{i=0}^{k} x_i}{\sum_{i=0}^{k} n_i} = \hat{p} \tag{4.48}$$

For the alternative model H_A^j, where j is the position of the Change-points, we calculate the estimator in two steps. In the first step, we estimate the values of partition sets before Change-point and the values after Change-point separately by Equation 4.27

$$\begin{aligned} \check{p}_{j,0} = ... = \check{p}_{j,j-1} &= \frac{\sum_{i=0}^{j-1} x_i}{\sum_{i=0}^{j-1} n_i} \\ \check{p}_{j,j} = ... = \check{p}_{j,k} &= \frac{\sum_{i=j}^{k} x_i}{\sum_{i=j}^{k} n_i} \end{aligned} \tag{4.49}$$

In the second step, we use the information from order restriction that the estimator before Change-point is smaller than the estimator after Change-point:$p_0 < p_k$. If $\check{p}_{j,0} \geq \check{p}_{j,k}$, a contradiction happens and the estimator should be calculated again to fulfill the requirement. Therefore, the MLE of H_A^j under order restriction should be calculated under two different situations. The first situation is that $\check{p}_{j,0} < \check{p}_{j,k}$, so we have two partition sets. The MLE of the lower level set is

$$\tilde{p}_{j,0} = ... = \tilde{p}_{j,j-1} = \check{p}_{j,0} \tag{4.50}$$

while the MLE of the higher level set is

$$\tilde{p}_{j,j} = ... = \tilde{p}_{j,k} = \check{p}_{j,k} \tag{4.51}$$

The second situation is that $\check{p}_0 \geq \check{p}_k$, so we have only one level set. All the estimators

are recalculated to avoid the conflicts and they achieve the maximum likelihood

$$\tilde{p}_{j,0} = ... = \tilde{p}_{j,k} = \frac{\sum_{i=0}^{k} x_i}{\sum_{i=0}^{k} n_i} = \hat{p} \tag{4.52}$$

Under the null hypothesis, all the parameters have the same mean and variance. The probability of the first situation is

$$P(\tilde{p}_{j,0} < \tilde{p}_{j,k}) = P(\frac{\sum_{i=0}^{j-1} x_i}{\sum_{i=0}^{j-1} n_i} < \frac{\sum_{i=j}^{k} x_i}{\sum_{i=j}^{k} n_i}) = 0.5 \tag{4.53}$$

Similarly, we can get the probability for the second situation as

$$P(\tilde{p}_{j,0} \geq \tilde{p}_{j,k}) = 0.5 \tag{4.54}$$

The AIC for model H_0 is

$$\begin{aligned} AIC(H_0) =& log(L(\hat{g}(x|\tilde{\theta_0}(x)))) - 1 \\ =& \{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\tilde{p}_{0,i}^{x_i} * (1 - \tilde{p}_{0,i})^{n_i - x_i})\} - 1 \end{aligned} \tag{4.55}$$

here $\tilde{\theta_0}(x) = \{\hat{p}, ..., \hat{p}\}$ is vector of the estimated parameters under null hypothesis H_0. The AIC for model H_A^j can be calculated as

$$\begin{aligned} AIC(H_A^j) =& log(L(\hat{g}(x|\tilde{\theta_j}(x)))) - r \\ =& \{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\tilde{p}_{j,i}^{x_i} * (1 - \tilde{p}_{j,i})^{n_i - x_i})\} - 2 \end{aligned} \tag{4.56}$$

here $\tilde{\theta_j}(x) = \{\tilde{p}_{j,0}, ..., \tilde{p}_{j,k}\}$ is vector of the MLE under different alternatives H_j. $r = 2$ is the number of unknown means.

In order to use the extra advantages of order restriction, Anraku (1999) introduced ORIC, which uses the one-sided information to calculate the penalty term under simple order restriction. We still note the method, which uses one-sided information and local MLE, as ORIC-lMLE. Under Single Change-point order restriction, the

ORIC-lMLE of different hypotheses are

$$\begin{aligned} ORIC-lMLE(H_0) =& log(L(\hat{g}(x|\tilde{\theta}_0(x)))) - 1 \\ =& \{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i-x_i)!}(\tilde{p}_{0,i}^{x_i} * (1-\tilde{p}_{0,i})^{n_i-x_i})\} - 1 \\ ORIC-lMLE(H_A^j) =& log(L(\hat{g}(x|\tilde{\theta}_j(x)))) - \sum_{\mathrm{m}=0}^{2} \mathrm{w}(2,\mathrm{m})\mathrm{m} \\ =& \{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i-x_i)!}(\tilde{p}_{j,i}^{x_i} * (1-\tilde{p}_{j,i})^{n_i-x_i})\} - 1.5 \end{aligned} \tag{4.57}$$

here

$$\begin{aligned} & \sum_{m=0}^{2} w(2,m)m \\ =& \sum_{l=0}^{2} l * P\{l,r,\omega\{H_A^j\}\} \\ =& 0 + 1 * 0.5 + 2 * 0.5 \\ =& 1.5 \end{aligned} \tag{4.58}$$

The example of adverse events rate: Change-point detection

Take the adverse events rate case given in previous section as an example. The researchers want to know if the adverse rate increases markedly at certain level of cabergoline. If the answer is yes, can this Change-point be estimated?

Treatment	Placebo	0.125(mg)	1.0(mg)
Present x_i	9	19	24
Absent $n_i - x_i$	11	24	17
Total n_i	20	43	41
$\hat{p_i}$	0.45	0.44	0.58

The hypotheses and estimators of partition sets are

Hypothesis	$\tilde{p_0}$	$\tilde{p_1}$	$\tilde{p_2}$	DF
H_0: $p_0 = p_1 = p_2$	$\frac{x_0+x_1+x_2}{n_0+n_1+n_2}$	$\frac{x_0+x_1+x_2}{n_0+n_1+n_2}$	$\frac{x_0+x_1+x_2}{n_0+n_1+n_2}$	1
H_A^1: $p_0 < p_1 = p_2$	$\frac{x_0}{n_0}$	$\frac{x_1+x_2}{n_1+n_2}$	$\frac{x_1+x_2}{n_1+n_2}$	2
H_A^2: $p_0 = p_1 < p_2$	$\frac{x_0+x_1}{n_0+n_1}$	$\frac{x_0+x_1}{n_0+n_1}$	$\frac{x_2}{n_2}$	2

By calculating the estimator of partition sets and adjusting contradictions by the following equation

$$\tilde{p}_{j,i} = \min_{l \geq i} \max_{m \leq i} \frac{\sum_{v=l}^{m} \check{p}_{j,v}}{\sum_{v=l}^{m} n_v} \tag{4.59}$$

we get the values of the local MLE as

Hypothesis	$\tilde{p}_{j,0}$	$\tilde{p}_{j,1}$	$\tilde{p}_{j,2}$
H_0: $p_0 = p_1 = p_2$	0.5000	0.5000	0.5000
H_A^1: $p_0 < p_1 = p_2$	0.4500	0.5119	0.5119
H_A^2: $p_0 = p_1 < p_2$	0.4444	0.4444	0.5854

Finally, we can calculate the log-likelihood, the ICs and the penalty terms by Equation 4.56 and 4.57 as following

Method	$H_0(T_{max})$	H_A^1	H_A^2	selected model
log-likelihood	-6.902	-6.779	-5.913	
ORIC-lMLE	-7.902	-8.279	-7.413	H_A^2
AIC	-7.902	-8.779	-7.913	H_A^2

here the penalty terms are given by Equation 4.58 as following

Penalty	H_0	H_A^1	H_A^2
ORIC-lMLE	1	1.5	1.5
AIC	1	2	2

From the result we can see that both AIC and ORIC-lMLE select model H_A^2 and draw the conclusion that higher dose of such medicine has the reverse effect. Since these two methods are model selection methods, neither of them can reject the null hypothesis with certain alpha level.

4.2.3 Estimators under Epidemic-order restriction

The MLE of H_0 is the same as former section. For the alternative model $H_A^{(s,j)}$, where r and s are the positions of the Change-points. The MLE of different partition sets are estimated separately

$$\begin{aligned}\check{p}_{j,0} = ... = \check{p}_{j,r-1} = \check{p}_{j,s} = ... = \check{p}_{j,k} &= \frac{\sum_{i=0}^{r-1} x_i + \sum_{i=s}^{k} x_i}{\sum_{i=0}^{r-1} n_i + \sum_{i=s}^{k} n_i} \\ \check{p}_{j,r} = ... = \check{p}_{j,s-1} &= \frac{\sum_{i=r}^{s-1} x_i}{\sum_{i=r}^{s-1} n_i}\end{aligned} \tag{4.60}$$

Similar as Single Change-point order restriction, we will use the information from order restriction that $p_r = p_s < p_0 = p_k$. If a contradiction happens, then the estimators should be calculated again to fulfill the requirement. So, the MLE for $H_A^{r,s}$ under order restriction has two situations. The first situation is that $\check{p}_{j,r} < \check{p}_{j,k}$, so we have two partition sets. The MLE of the higher level set is

$$\tilde{p}_{j,0} = ... = \tilde{p}_{j,r-1} = \tilde{p}_{j,s} = ... = \tilde{p}_{j,k} = \check{p}_{j,0} = \frac{\sum_{i=0}^{r-1} x_i + \sum_{i=s}^{k} x_i}{\sum_{i=0}^{r-1} n_i + \sum_{i=s}^{k} n_i} \tag{4.61}$$

and the MLE of the lower level set between the Change-points is

$$\tilde{p}_{j,r} = ... = \tilde{p}_{j,s-1} = \check{p}_{j,r} = \frac{\sum_{i=r}^{s-1} x_i}{\sum_{i=r}^{s-1} n_i} \tag{4.62}$$

and the second situation is that $\check{p}_{j,0} \geq \check{p}_{j,k}$, so we have only one partition sets. All the estimators are the same

$$\tilde{p}_{j,0} = ... = \tilde{p}_{j,k} = \frac{\sum_{i=0}^{k} x_i}{\sum_{i=0}^{k} n_i} = \hat{p} \tag{4.63}$$

Under the null hypothesis, the probability of the first situation is

$$P(\check{p}_{j,0} < \check{p}_{j,r}) = 0.5 \tag{4.64}$$

Similarly, we get the probability for the second situation as

$$P(\check{p}_{j,0} \geq \check{p}_{j,r}) = 0.5 \tag{4.65}$$

For multiple Change-points problem the MLE can be calculated in the similar way. The parameters are divided by Change-points into different partition sets. The pooled mean can be taken as the MLE. Similarly we can calculate the AIC and ORIC-lMLE.

Under epidemic Change-points order restriction, the NIC for different hypotheses are calculated from Equation 4.32 and 4.33 as

$$\begin{aligned} NIC(H_0) =& log(L(\hat{g}(x|\tilde{\theta}(x)))) - 1 \\ =& -\{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\tilde{p}_{j,i}^{x_i} * (1 - \tilde{p}_{j,i})^{n_i - x_i})\} - 1 \end{aligned} \tag{4.66}$$

and

$$\begin{aligned} NIC(H_A^j) =& log(L(\hat{g}(x|\tilde{\theta}_j(x)))) - 2 - 3 * m \\ =& -\{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\tilde{p}_{j,i}^{x_i} * (1 - \tilde{p}_{j,i})^{n_i - x_i})\} - 2 - \mathrm{b}(\widetilde{\mathrm{p}}^{(1)}, \widetilde{\mathrm{p}}^{(2)}) - \mathrm{b}(\widetilde{\mathrm{p}}^{(2)}, \widetilde{\mathrm{p}}^{(1)}) \end{aligned} \tag{4.67}$$

DNA-motif finding

In previous section, we have successfully transformed the DNA-motif finding problem into a contingency table (Table 4.2). The aim for this study is to find out the Change-points and to locate the motif around Change-points. The MLEs are calculated similarly as Single Change-point order restriction. The log-likelihood is calculated with given MLE and the penalty term is equal to one under null hypothesis and 8

Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
x_i	14	14	13	7	9	6	9	8	8	6	10	7	6	8	12	14	14
n_i	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
$\hat{p}_i$	1	1	.9	.5	.6	.4	.6	.6	.6	.4	.7	.5	.4	.6	.9	1	1

Table 4.2: Contingency table for the DNA-motif

under alternative hypothesis. The second penalty term is the sum of penalty from two Change-points and two unknown parameters (Ninomiya, 2005). Since we have two unknown Change-points, the results are listed in a two-dimension table. Here is part of the table

	s=13	s=14	s=15	s=16
r=1	-51.2	-44.9	-50.5	-57.1
r=2	-45.4	-38.1	-43.3	-50.5
r=3	-40.6	**-33.4**	-35.1	-42.3
r=4	-46.7	-42.7	-45.4	-51.2

and the value of the null model is -54.1. We can select position "3" and "14" as the best prediction of the Change-points.

4.2.4 Estimators under Simple-order restriction

The Simple-order restriction can be treated as multiple Change-points problem which has ordered means. The inequality relationship is considered as ordered Change-points. If two near partition sets are in the second situation that the order restriction does not hold, we can take the pooled mean as the average mean for both of them. Many model selection methods, such as AIC and ORIC-lMLE, are available.

Under Simple-order restriction, the ORIC-lMLE of different hypotheses are

$$\begin{aligned} ORIC-lMLE(H_0) =& log(L(\hat{g}(x|\tilde{\theta_0}(x)))) - 1 \\ =& -\{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i-x_i)!}(\tilde{p}_{0,i}^{x_i} * (1-\tilde{p}_{0,i})^{n_i-x_i})\} - 1 \\ ORIC-lMLE(H_A^j) =& log(L(\hat{g}(x|\tilde{\theta_j}(x)))) - \sum_{m=0}^{k} w(k,m)m \\ =& -\{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i-x_i)!}(\tilde{p}_{j,i}^{x_i} * (1-\tilde{p}_{j,i})^{n_i-x_i})\} - \sum_{l=1}^{r} l * P\{l,r,\omega\{H_A^j\}\} \end{aligned} \tag{4.68}$$

here, l is the number of partition sets under H_A^j and $P\{l,r,\omega\{H_A^j\}\}$ is the level probability defined by Robertson et al. (1988).

However, ORIC-lMLE does not work well in selecting the correct model because of the "underperformance", which we have discussed in former section. Later, the simulation study also verifies this conclusion. We have developed a new IC, which use SLE and called Mi and Hothorn IC (MHIC), to solve this problem.

Our new IC (MHIC) are defined as

$$\begin{aligned} MHIC(H_0) = AIC(H_0) =& log(L(\hat{g}(x)|\bar{\theta_0}(x))) - 1 \\ =& -\{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i-x_i)!}(\bar{p}_{0,i}^{x_i} * (1-\bar{p}_{0,i})^{n_i-x_i})\} - 1 \\ MHIC(H_A^j) =& log(L(\hat{g}(x)|\bar{\theta_j}(x))) - \sum_{m=0}^{2} w(2,m)m \\ =& -\{\sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i-x_i)!}(\bar{p}_{j,i}^{x_i} * (1-\bar{p}_{j,i})^{n_i-x_i})\} - 1.5 \end{aligned} \tag{4.69}$$

Males age	< 25	$25-29$	$30-34$	$35-39$
Abortion x_i	33	37	3	7
Normal $n_i - x_i$	226	321	358	5
Total n_i	259	358	64	12
$\hat{p}_i$	0.127	0.103	0.047	0.583

Table 4.3: Spontaneous abortion rate.

here

$$\begin{aligned}&\sum_{m=0}^{2} w(2,m)m\\ =&\sum_{l=0}^{2} l * P\{l,r,\omega\{H_A^j\}\\ =&0+1*0.5+2*0.5=1.5\end{aligned} \tag{4.70}$$

$\bar{\theta}_0(x) = \{\bar{p},...,\bar{p}\}$ is the vector of the estimated parameters under null hypothesis H_0 and $\bar{\theta}_j(x) = \{\bar{p}_{j,0},...,\bar{p}_{j,k}\}$ is the SLE vector under alternatives H_A^j. From the following example, we will see that the MHIC with SLE achieves worse KL distance than ORIC-lMLE for the wrong model. MHIC has stronger power to exclude the wrong model.

Spontaneous abortion rate

The hypotheses and estimators of partition sets are

Hypothesis	$\breve{p}_{j,0}$	$\breve{p}_{j,1}$	$\breve{p}_{j,2}$	$\breve{p}_{j,3}$	DF
H_0: $p_0 = p_1 = p_2 = p_3$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	1
H_A^1: $p_0 < p_1 = p_2 = p_3$	$\frac{x_0}{n_0}$	$\frac{x_1+x_2+x_3}{n_1+n_2+n_3}$	$\frac{x_1+x_2+x_3}{n_1+n_2+n_3}$	$\frac{x_1+x_2+x_3}{n_1+n_2+n_3}$	2
H_A^2: $p_0 = p_1 < p_2 = p_3$	$\frac{x_0+x_1}{n_0+n_1}$	$\frac{x_0+x_1}{n_0+n_1}$	$\frac{x_2+x_3}{n_2+n_3}$	$\frac{x_2+x_3}{n_2+n_3}$	2
H_A^3: $p_0 = p_1 = p_2 < p_3$	$\frac{x_0+x_1+x_2}{n_0+n_1+n_2}$	$\frac{x_0+x_1+x_2}{n_0+n_1+n_2}$	$\frac{x_0+x_1+x_2}{n_0+n_1+n_2}$	$\frac{x_3}{n_3}$	2
H_A^4 : $p_0 < p_1 < p_2 < p_3$	$\frac{x_0}{n_0}$	$\frac{x_1}{n_1}$	$\frac{x_2}{n_2}$	$\frac{x_3}{n_3}$	4
H_A^5 : $p_0 = p_1 < p_2 < p_3$	$\frac{x_0+x_1}{n_0+n_1}$	$\frac{x_0+x_1}{n_0+n_1}$	$\frac{x_2}{n_2}$	$\frac{x_3}{n_3}$	3
H_A^6 : $p_0 < p_1 = p_2 < p_3$	$\frac{x_0}{n_0}$	$\frac{x_1+x_2}{n_1+n_2}$	$\frac{x_2}{n_2}$	$\frac{x_3}{n_3}$	3
H_A^7 : $p_0 < p_1 < p_2 = p_3$	$\frac{x_0}{n_0}$	$\frac{x_1}{n_1}$	$\frac{x_2}{n_2}$	$\frac{x_3}{n_3}$	3

here $\hat{p} = \frac{x_0+x_1+x_2+n_3}{n_0+n_1+n_2+n_3}$ is the overall mean. By using the following equation

$$\tilde{p}_{j,i} = \min_{l \geq i} \max_{m \leq i} \frac{\sum_{l}^{v=m} \check{p}_{j,v}}{\sum_{l}^{v=m} n_v} \tag{4.71}$$

the values of local MLE are calculated as

Hypothesis	$\tilde{p}_{j,0}$	$\tilde{p}_{j,1}$	$\tilde{p}_{j,2}$	$\tilde{p}_{j,3}$
H_0: $p_0 = p_1 = p_2 = p_3$	0.115	0.115	0.115	0.115
H_A^1: $p_0 < p_1 = p_2 = p_3$	0.115	0.115	0.115	0.115
H_A^2: $p_0 = p_1 < p_2 = p_3$	0.113	0.113	0.131	0.131
H_A^3: $p_0 = p_1 = p_2 < p_3$	0.107	0.107	0.107	0.583
H_A^4 : $p_0 < p_1 < p_2 < p_3$	0.107	0.107	0.107	0.583
H_A^5 : $p_0 = p_1 < p_2 < p_3$	0.107	0.107	0.107	0.583
H_A^6 : $p_0 < p_1 = p_2 < p_3$	0.107	0.107	0.107	0.583
H_A^7 : $p_0 < p_1 < p_2 = p_3$	0.113	0.113	0.131	0.131

while the SLE are

Hypothesis	$\hat{p}_{j,0}$	$\hat{p}_{j,1}$	$\hat{p}_{j,2}$	$\hat{p}_{j,3}$
H_0: $p_0 = p_1 = p_2 = p_3$	0.115	0.115	0.115	0.115
H_A^1: $p_0 < p_1 = p_2 = p_3$	0.115	0.115	0.115	0.115
H_A^2: $p_0 = p_1 < p_2 = p_3$	0.113	0.113	0.131	0.131
H_A^3: $p_0 = p_1 = p_2 < p_3$	0.107	0.107	0.107	0.583
H_A^4 : $p_0 < p_1 < p_2 < p_3$	0.077	0.102	0.128	0.153
H_A^5 : $p_0 = p_1 < p_2 < p_3$	0.090	0.090	0.115	0.165
H_A^6 : $p_0 < p_1 = p_2 < p_3$	0.065	0.115	0.115	0.165
H_A^7 : $p_0 < p_1 < p_2 = p_3$	0.065	0.115	0.140	0.140

which are calculated by the following equations

Penalty	H_0	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5	H_A^6	H_A^7
ORIC-lMLE	1	1.5	1.5	1.5	2.07	1.83	1.83	1.83
MHIC	1	1.5	1.5	1.5	1.5	1.5	1.5	1.5

Table 4.4: Penalties of the ICs

Hypothesis	$\hat{p}_{j,0}$	$\hat{p}_{j,1}$	$\hat{p}_{j,2}$	$\hat{p}_{j,3}$	DF
H_0: $p_0 = p_1 = p_2 = p_3$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	1
H_A^1: $p_0 < p_1 = p_2 = p_3$	$\hat{p} + \frac{c_{1,0}}{C_1} * \Delta$	$\hat{p} + \frac{c_{1,1}}{C_1} * \Delta$	$\hat{p} + \frac{c_{1,2}}{C_1} * \Delta$	$\hat{p} + \frac{c_{1,3}}{C_1} * \Delta$	2
H_A^2: $p_0 = p_1 < p_2 = p_3$	$\hat{p} + \frac{c_{2,0}}{C_2} * \Delta$	$\hat{p} + \frac{c_{2,1}}{C_2} * \Delta$	$\hat{p} + \frac{c_{2,2}}{C_1} * \Delta$	...	2
H_A^3: $p_0 = p_1 = p_2 < p_3$	$\hat{p} + \frac{c_{3,0}}{C_3} * \Delta$	$\hat{p} + \frac{c_{3,1}}{C_3} * \Delta$	...	...	2
H_A^4 : $p_0 < p_1 < p_2 < p_3$	$\hat{p} + \frac{c_{4,0}}{C_4} * \Delta$	...	...	...	2
H_A^5 : $p_0 = p_1 < p_2 < p_3$	$\hat{p} + \frac{c_{5,0}}{C_5} * \Delta$	...	...	...	2
H_A^6 : $p_0 < p_1 = p_2 < p_3$	$\hat{p} + \frac{c_{6,0}}{C_6} * \Delta$	...	...	...	2
H_A^7 : $p_0 < p_1 < p_2 = p_3$	$\hat{p} + \frac{c_{7,0}}{C_7} * \Delta$	...	...	...	2

The contrasts of MCT for different hypothesis are

Hypothesis	$c_{j,0}$	$c_{j,1}$	$c_{j,2}$	$c_{j,3}$
H_A^1: $p_0 < p_1 = p_2 = p_3$	$c_{1,0} = -3$	$c_{1,1} = 1$	$c_{1,2} = 1$	$c_{1,3} = 1$
H_A^2: $p_0 = p_1 < p_2 = p_3$	$c_{2,0} = -2$	$c_{2,1} = -2$	$c_{2,2} = 2$	$c_{2,3} = 2$
H_A^3: $p_0 = p_1 = p_2 < p_3$	$c_{3,0} = -1$	$c_{3,1} = -1$	$c_{3,2} = -1$	$c_{3,3} = 3$
H_A^4 : $p_0 < p_1 < p_2 < p_3$	$c_{4,0} = -3$	$c_{4,1} = -1$	$c_{4,2} = 1$	$c_{4,3} = 3$
H_A^5 : $p_0 = p_1 < p_2 < p_3$	$c_{5,0} = -1$	$c_{5,1} = -1$	$c_{5,2} = 0$	$c_{5,3} = 2$
H_A^6 : $p_0 < p_1 = p_2 < p_3$	$c_{6,0} = -1$	$c_{4,1} = 0$	$c_{4,2} = 0$	$c_{4,3} = 1$
H_A^7 : $p_0 < p_1 < p_2 = p_3$	$c_{7,0} = -2$	$c_{4,1} = 0$	$c_{4,2} = 1$	$c_{4,3} = 1$

The final ICs and penalty term are given as following by Equation 4.68 and 4.69

Meth.	$H_0(T_{max})$	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5	H_A^6	H_A^7	Sel.M.
MHIC	-19.08	-19.58	-19.48	-11.78	-21.74	-19.21	-23.73	-25.90	H_A^3
ORIC-lMLE	-19.08	-19.58	-19.48	-11.78	-12.35	-12.11	-12.11	-19.81	H_A^3

The penalties are listed in Table 4.4 here the penalty term of MHIC is calculated by Equation 4.69. We just give the detail for the penalty term of ORIC-lMLE which is

calculated by Equation 4.68. The values under model H_A^1, H_A^2 and H_A^3 are similar as what we have before and easy to be calculated as

$$\begin{aligned}
&\sum_{l=1}^{r} l * P\{l, r, \omega\{H_A^1\} \\
=&\sum_{l=1}^{2} l * P\{l, r, \omega\{H_A^1\} \\
=&1 * 0.5 + 2 * 0.5 = 1.5
\end{aligned} \tag{4.72}$$

The value under model H_A^5, H_A^6, H_A^7 are similar. So here we calculate only for model H_5, which have $r = 3$ partition sets. The probability is

$$\begin{aligned}
&\sum_{l=1}^{r} l * P\{l, r, \omega\{H_A^5\}\} \\
=&\sum_{l=1}^{3} l * P\{l, 3, \omega\{H_A^5\}\} \\
=&1 * P\{1, 3, \omega\{H_A^5\}\} + 2 * P\{2, 3, \omega\{H_A^5\}\} + 3 * P\{3, 3, \omega\{H_A^5\}\} \\
=&1.83
\end{aligned} \tag{4.73}$$

The value under model H_4 has $r = 4$ partition sets. The probability is

$$\begin{aligned}
&\sum_{l=1}^{r} l * P\{l, r, \omega\{H_A^4\} \\
=&\sum_{l=1}^{4} l * P\{l, 4, \omega\{H_A^4\} \\
=&1 * P\{1, 4, \omega\{H_A^4\}\} + 2 * P\{2, 4, \omega\{H_A^4\}\} + 3 * P\{3, 4, \omega\{H_A^4\}\} + 4 * P\{4, 4, \omega\{H_A^4\}\} \\
=&2.07
\end{aligned} \tag{4.74}$$

The conclusion for the problem is that we select model H_j^3 as the best model. Because both MHIC and ORIC-lMLE methods are model selection methods, we cannot control the FWER. These two methods are not identical under Simple order restriction. For ORIC-lMLE method, the degrees of freedom for the chi-square distribution

are larger than two.

4.2.5 Estimators under Simple-tree order restriction

The hypothesis under Simple-tree order restriction can be described as

$$\begin{aligned} H_0 &: p_0 = p_1 = ... = p_k \\ H_A &: \bigcup_{j=1}^{k} p_0 < p_j \end{aligned} \tag{4.75}$$

Each elementary alternative model has only two parameters while the global null hypotheses have $k+1$ parameters. In order to make the ratio test, we separate the null hypothesis into elementary null models too.

$$\begin{aligned} H_0 &: \bigcup_{j=1}^{k} p_0 = p_j \\ H_A &: \bigcup_{j=1}^{k} p_0 < p_j \end{aligned} \tag{4.76}$$

We calculate the difference of log-likelihood directly from this two elementary models.

$$ratio = L(H_A^j) - L(H_0^j) \tag{4.77}$$

The MLE here is different as what we have before. For H_0^j

$$\hat{p}_{0,0} = \tilde{p}_{0,0} = \hat{p}_{j,0} = \tilde{p}_{j,0} = \frac{x_0 + x_j}{n_0 + n_j} \tag{4.78}$$

For H_A^j

$$\hat{p}_{j,0} = \tilde{p}_{j,0} = min(\frac{x_0}{n_0}, \frac{x_0 + x_i}{n_0 + n_i}) \tag{4.79}$$

and

$$\hat{p}_{j,j} = \tilde{p}_{j,j} = max(\frac{x_j}{n_j}, \frac{x_0 + x_j}{n_0 + n_j}) \tag{4.80}$$

Here we use OSAIC to calculate the ratio too, the equation is

$$OSAIC = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\hat{p}_{j,i}^{x_i} * (1 - \hat{p}_{j,i})^{n_i - x_i}) - \sum_{m=0}^{r} w(r,m)(k - r + m) \tag{4.81}$$

for the elementary null models, there is no inequality, so $r = 0$

$$OSAIC(H_0^j) = \sum_{i=0,j} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\hat{p}_{i,0}^{x_i} * (1 - \hat{p}_{i,0})^{n_i - x_i}) - 1 \tag{4.82}$$

since we have only inequality for the elementary alternative models i.e. $r = 1$ implies

$$OSAIC(H_A^j) = \sum_{i=0,j} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\hat{p}_{j,i}^{x_i} * (1 - \hat{p}_{j,i})^{n_i - x_i}) - 1.5 \tag{4.83}$$

the ratio is $OSAIC(H_A^j) - OSAIC(H_0^j)$

The example of adverse events rate: many-to-one comparison

Take the adverse events rate case given in previous section as an example. The researchers want to know if the dose groups are significantly different to the control group. If the answer is yes, can these groups be estimated?

Treatment	Placebo	0.125(mg)	1.0(mg)
Present x_i	9	19	24
Absent $n_i - x_i$	11	24	17
Total n_i	20	43	41
$\hat{p_i}$	0.45	0.44	0.58

The hypotheses and estimators of partition sets are

Hypothesis	$\check{p_0}$	$\check{p_1}$	$\check{p_2}$	DF
H_0: $p_0 = p_1 = p_2$	$\frac{x_0+x_1+x_2}{n_0+n_1+n_2}$	$\frac{x_0+x_1+x_2}{n_0+n_1+n_2}$	$\frac{x_0+x_1+x_2}{n_0+n_1+n_2}$	1
H_A^1: $p_0 < p_1$	$\frac{x_0}{n_0}$	$\frac{x_1}{n_1}$		1
H_A^2: $p_0 < p_2$	$\frac{x_0}{n_0}$		$\frac{x_2}{n_2}$	1

By calculating the estimator of partition sets and adjusting contradictions by the following equations

For H_0^j

$$\hat{p}_{0,0} = \tilde{p}_{0,0} = \hat{p}_{j,0} = \tilde{p}_{j,0} = \frac{x_0 + x_j}{n_0 + n_j} \tag{4.84}$$

For H_A^j

$$\hat{p}_{j,0} = \tilde{p}_{j,0} = min(\frac{x_0}{n_0}, \frac{x_0 + x_i}{n_0 + n_i}) \tag{4.85}$$

and

$$\hat{p}_{j,j} = \tilde{p}_{j,j} = max(\frac{x_j}{n_j}, \frac{x_0 + x_j}{n_0 + n_j}) \tag{4.86}$$

we get the values of the local MLE as

Hypothesis	$\tilde{p}_{j,0}$	$\tilde{p}_{j,1}$	$\tilde{p}_{j,2}$
H_0^1: $p_0 = p_1$	0.445	0.445	
H_A^1: $p_0 < p_1$	0.45	0.45	
H_0^2: $p_0 = p_2$	0.515		0.515
H_A^2: $p_0 < p_2$	0.45		0.58

Finally, we can calculate the log-likelihood, the ICs and the penalty terms by Equation as following

Method	$H_0(T_{max})$	H_A^1	H_A^2	selected model
ORIC-lMLE	-0	-0.500	-0.004	H_0

here the penalty terms are given by Equation 4.58 too, as following

Penalty	H_0	H_A^1	H_A^2
ORIC-lMLE	1	1.5	1.5

From the result, we can see that ORIC-lMLE selects model H_0. We can draw the conclusion that no dose of such medicine is significantly different to the control. This result is different from what we get before. The ICs under Simple-tree order restrictions has less information than under Simple-order.

Since this method is model selection method, it cannot reject the null hypothesis with certain α level either.

4.3 Simulation study for comparing gMLE, lMLE and SLE

In last section we have introduced gMLE, lMLE and SLE to calculate the likelihood. Theoretically, they can be interpreted as projections (estimations) to different spaces. This influences the final classification rate. In this section, we will give a short simulation example of the correct model selection rate over different alternatives under Simple-order restriction. In these simulations, we generate 1000 random binomial data for $k = 3$ isotonic means $0.4 \leq p_0 \leq ... \leq p_k \leq 0.4 + \Delta = 0.6$, and

Alternatives	Methods	$H_0(Theor.)$	H_0	H_A^1	H_A^2	H_A^3
0.4/0.6/0.6	ORIC-gMLE	-	0.004	0.555	0.014	0.427
0.4/0.6/0.6	ORIC-lMLE	-	0.001	0.568	0.013	0.418
0.4/0.6/0.6	MHIC-SLE	0.0073	0.002	0.811	0.009	0.178
0.4/0.4/0.6	ORIC-gMLE	-	0.008	0.014	0.557	0.422
0.4/0.4/0.6	ORIC-lMLE	-	0.006	0.013	0.586	0.395
0.4/0.4/0.6	MHIC-SLE	0.0073	0.007	0.010	0.817	0.166
0.4/0.5/0.6	ORIC-gMLE	-	0.021	0.223	0.221	0.536
0.4/0.5/0.6	ORIC-lMLE	-	0.020	0.221	0.233	0.526
0.4/0.5/0.6	MHIC-SLE	0.0170	0.020	0.236	0.224	0.520

Table 4.5: 1000 random binomial data for $k = 3$, proportions $p_0 = ... = p_{j-1} = 0.4$,$p_j = 0.4, 0.5, 0.6$, $p_{j+1} = ... = p_k = 0.6$, and sample size n_i is 100.

Method	IC	New IC?	Estimator	New Estimator?
ORIC-gMLE	ORIC	NO	gMLE	NO
ORIC-lMLE	ORIC	NO	lMLE	**YES**
MHIC-SLE	MHIC	**YES**	SLE	**YES**

Table 4.6: Our NEW IC with NEW estimator

the sample size is 100. ORIC with lMLE, ORIC with gMLE and our new method MHIC which uses SLE, are compared together.

From this simulation study, we see that ORIC with lMLE is slightly better than ORIC with gMLE. However, both of them have a very high misclassification rate in identifying model H_A^1 and H_A^2 under Simple-order restriction.

An improvement for ORIC with lMLE could be made. We leave the question of how to calculate a suitable penalty for ORIC with lMLE as an open question.

Chapter 5

Test-based model selection

In the last chapter we have already reviewed former model-based methods and introduced our MHIC method for solving the previous problems. Simulation studies in Chapter 6.2 shows that MHIC with "suitable" likelihood estimators (SLE), achieves a higher model selection rate than traditional ORIC method. However, model-based methods are designed for model selection, not test. They cannot control the FWER. In this chapter, we will introduce a test-based method, which uses the IC value developed from last chapter as test statistics and controls the FWER. We will also study the "power", correct model selection rate (CR) and misclassification rate (MR) for all these methods. First, we will discuss about the distribution of the likelihood. Second, the critical value will be calculated from the quantile of certain distribution to control the FWER.

5.1 Relationship between Log-likelihood Ratio Test (LRT) and Multiple Contrast Test (MCT)

In this section, we will compare the relationship between LRT and MCT. Because MCT is a 'collection' of Single Contrast Test (SCT), we first figure out the relationship between SCT and LRT.

For SCT, the choice of the contrast coefficients c_is is free. According to Wright (1988), a SCT is very powerful if we have some previous knowledge of the true mean vector $U = \{u_0, u_1, ..., u_k\}$. However, in order to achieve a good power in "average" as LRT, they suggested choose c_is properly such that c_is are "strategically" located in the whole alternative space. Here we describe the idea of them briefly. The alternative space of LRT is a polyhedral cone and the null hypothesis is a linear subspace on the boundary this cone (Pincus, 1975). For different order restrictions, the region of the polyhedral cone is different. For Simple-tree order restriction, Wright (1988) developed corner vectors which are the edges of the cone. Furthermore, they suggested people using orthogonal contrasts, which are linear combinations of the corner vector, for simplifying the calculation. Contrasts with similar functionality are developed by Hirotsu and Marumo (2002) for Single Change-point order restriction and simple-order restriction. "Then maximal contrast type test is derived systematically as the likelihood ratio test for each of those Change-point hypotheses" (Hirotsu and Marumo, 2002).

5.1.1 Distribution of the log-likelihood under Single Change-point order restriction

According to Robertson et al. (1988), for normal distributed data under the null hypothesis, the difference of log-likelihood between the alternative model and the null model are distributed as a weighted chi-squared distribution.

The critical value of weighted chi-squared distribution is hard to calculate. But we can transform the distribution into square of normal distribution if the data are under Single Change-point order restriction. We have $q = k$ different elementary alternatives.

In the following prove, we assume that the sample size is large enough, so the binomial data can be treated asymptotically as normal distributed. The distribution of

the differences for elementary alternatives H_A^j against the H_0 are noted as TC_j that

$$\begin{aligned} TC_j &= \log(L(\hat{g}(\hat{\theta}_j|x)), x \in H_A^j) - \log(L(\hat{g}(\hat{\theta}_0|x)), x \in H_0) \\ &= \{-0.5N\log(2\pi) - 0.5\sum_0^k n_i \log\hat{\sigma}_i - 0.5\sum_0^k \omega_i(x_i - \hat{p}_{j,i})^2\} \\ &- \{-0.5N\log(2\pi) - 0.5\sum_0^k n_i \log\hat{\sigma} - 0.5\sum_0^k \omega_i(x_i - \hat{p})^2\} \end{aligned} \tag{5.1}$$

We can omit all the terms that p_is are not involved (Robertson et al., 1988; Hughes and King, 2003).

$$\begin{aligned} TC_j &\approx -0.5\{\sum_0^k \omega_i(x_i - \hat{p}_{j,i})^2 - \sum_0^k \omega_i(x_i - \hat{p})^2\} \\ &= -0.5\{(\sum_0^k \omega_i(\hat{p}_{j,i} - \hat{p})^2 + \underbrace{2\sum_0^k \omega_i(x_i - \hat{p})(\hat{p} - \hat{p}_{j,i}))}_{=0}\} \\ &= -0.5\sum_0^k \omega_i(\hat{p}_{j,i} - \hat{p})^2 \\ &\sim -0.5\chi_{01}^2 \end{aligned} \tag{5.2}$$

here, $\tilde{\theta}_j = \{\tilde{p}_{j,0}, ..., \tilde{p}_{j,k}\}$ is the vector of the estimators for elementary alternative H_A^j, $\tilde{\theta}_0 = \{\hat{p}, ..., \hat{p}\}$ is the vector of the estimators for null hypothesis H_0 and $\omega_i = n_i/\hat{\sigma_i^2}$, $\sum_0^k \omega_i = 1$ and $\hat{\sigma_i^2}$ is the estimated variances. The distribution of χ_{01}^2 is,

$$P[0.5\chi_{01}^2 \geq c] = \sum_{m=1}^{l} P\{m, l, \omega\{H_i\}\}P[0.5\chi_{df=m-1}^2 \geq c] \tag{5.3}$$

Here, $\chi_{df=m-1}^2$ is univariate chi-square distribution with df degrees of freedom and $\chi_0^2 = 0$, l is the number of total partition sets under hypothesis and $P\{m, l, \omega\{H_i\}\}$ is the level probability that under hypothesis $\{H_i\}$, there are m different values among the l partition sets. If we have m different values, then the number of Change-point

is $m-1$. (Robertson et al., 1988; Hughes and King, 2003)

For the special case of Single Change-point order restriction, the level probability is quite simple

$$P\{1,2,\omega\{H_i\}\} = P\{2,2,\omega\{H_i\}\} = 0.5 \tag{5.4}$$

the critical value for given α is (Xiong and Barmi, 2002)

$$\begin{aligned}\alpha = P\{0.5\chi^2_{01} \geq z_{1-\alpha}\} =& 0.5P\{0.5\chi^2_{df=1} \geq z_{1-\alpha}\} + 0.5\underbrace{P\{0.5\chi^2_{df=0} \geq z_{1-\alpha}\}}_{0} \\ =& 0.5P\{0.5\chi^2_{df=1} \geq z_{1-\alpha}\}\end{aligned} \tag{5.5}$$

then we have the following relationship

$$P\{0.5\chi^2_{df=1} \geq z_{1-\alpha}\} = 2\alpha \tag{5.6}$$

The critical value for single elementary alternative can be calculated by a one-sided chi-square distribution

$$z_{1-\alpha} = 0.5\chi^2_{df=1}\{p = (1-2\alpha)\} \tag{5.7}$$

The log-likelihood ratios of all hypotheses minus the null are multivariate weighted chi-square distributed with covariance matrix Σ_{chi}.

Xiong and El Barmi (2002) indicated that the distribution is complicated and gave a simulated critical value sc_α

$$P\{[\max_{1\leq j\leq k} \log(L(\hat{g}(x), x \in H_A^j)) - \log(L(\hat{g}(x), x \in H_0))] \geq sc_\alpha\} = \alpha \tag{5.8}$$

However, for Single Change-point problem, the critical value can be calculated by multivariate normal distribution. Here we first assume balanced design that $n_0 = ... = n_k = n$. By extending the MLE into Equation 5.2, we get the value of new test statistics TC_j for Single Change-point problem.

As we have discussed in last chapter, when $\check{p_0} < \check{p_k}$, there are two partition sets. The probability for this event is $P\{2, 2, \omega\{H_i\}\} = 0.5$. The MLEs are

$$\begin{aligned}
\tilde{p}_{j,0} = ... = \tilde{p}_{j,j-1} &= \frac{\sum_{i=0}^{j-1} x_i}{\sum_{i=0}^{j-1} n_i} = \frac{\sum_{i=0}^{j-1} x_i}{jn} \\
\tilde{p}_{j,j} = ... = \tilde{p}_{j,k} &= \frac{\sum_{i=j}^{k} x_i}{\sum_{i=j}^{k} n_i} = \frac{\sum_{i=j}^{k} x_i}{(k-j+1)n} \\
\hat{p} &= \frac{\sum_{i=0}^{k} x_i}{\sum_{i=0}^{k} n_i} = \frac{\sum_{i=0}^{k} x_i}{(k+1)n}
\end{aligned} \tag{5.9}$$

By putting above values in Equation 5.2, we have

$$\begin{aligned}
&TC_j = 0.5 \sum_{0}^{k} \omega_i (\tilde{p}_{j,i} - \hat{p})^2 \\
=&0.5\omega\{j(\frac{\sum_{i=0}^{j-1} x_i}{jn} - \frac{\sum_{i=0}^{k} x_i}{n(k+1)})^2 + (k-j+1)(\frac{\sum_{i=j}^{k} x_i}{(k-j+1)n} - \frac{\sum_{i=0}^{k} x_i}{n(k+1)})^2\} \\
=&0.5\omega\{j(\frac{(k+1)\sum_{i=0}^{j-1} x_i}{jn(k+1)} - \frac{j\sum_{i=0}^{k} x_i}{jn(k+1)})^2 \\
&+ (k-j+1)(\frac{(k+1)\sum_{i=j}^{k} x_i}{(k-j+1)n(k+1)} - \frac{(k-j+1)\sum_{i=0}^{k} x_i}{(k-j+1)n(k+1)})^2\} \\
=&0.5\frac{n}{\sigma^2}\{j(\frac{(k-j+1)\sum_{i=0}^{j-1} x_i}{jn(k+1)} - \frac{j\sum_{i=j}^{k} x_i}{jn(k+1)})^2 \\
&+ (k-j+1)(\frac{j\sum_{i=j}^{k} x_i}{(k-j+1)n(k+1)} - \frac{(k-j+1)\sum_{i=0}^{j-1} x_i}{(k-j+1)n(k+1)})^2\} \\
=&0.5\frac{n}{\sigma^2}\{\sum_{i=0}^{j-1} -(k-j+1)x_i + \sum_{i=j}^{k} jx_i\}^2 * \{j(\frac{1}{jn(k+1)})^2 + (k-j+1)(\frac{1}{(k-j+1)n(k+1)})^2\} \\
=&0.5\frac{n}{\sigma^2}\{\sum_{i=0}^{j-1} c_{j,i}x_i + \sum_{i=j}^{k} c_{j,i}x_i\}^2 * \{\frac{1}{j(k-j+1)(k+1)n^2}\} \\
=&0.5\{\frac{\sum_{i=0}^{k} c_{j,i}x_i}{\sqrt{n\sigma^2(k+1)j(k-j+1)}}\}^2 \\
=&0.5\{\frac{\sum_{i=0}^{k} c_{j,i}x_i}{\sqrt{n\sigma^2 j(k-j+1)^2 + (k-j+1)j^2}}\}^2 \\
=&0.5\{\frac{\sum_{i=0}^{k} c_{j,i}x_i}{\sqrt{n\sigma^2 \sum_{i=0}^{j-1} c_{j,i} + \sum_{i=j}^{k} c_{j,i}^2}}\}^2 \\
=&0.5T_j^2
\end{aligned} \tag{5.10}$$

T_j is the test statistics of MCT for the same alternative H_A^j in Equation 3.22 and c_i is the contrasts for Change-point. We have proved that the value of TC_j is proportion to the square of T_j. When $\check{p_0} \geq \check{p_k}$, there is only one level set. All the MLEs have the same value that $\tilde{p_j} = \hat{p}$. The probability for this event is $P\{1, 2, \omega\{H_i\}\} = 0.5$. Put values above in Equation 5.2, we have

$$\begin{aligned} & TC_j \\ =& 0.5 \sum_0^k \omega_i (\tilde{p}_{j,i} - \hat{p})^2 \\ =& 0 \end{aligned} \tag{5.11}$$

Simply to prove, in this case the correspond MCT $T_j < 0$. For Single Change-point problem, the critical value ($c_\alpha > 0$) of multivariate weighted chi-square distribution can be calculated from multi normal distribution

$$\begin{aligned} \alpha =& P(\max_{1\leq j\leq k} 2TC_j > 2z_{1-\alpha}) \\ =& \underbrace{P\{1,2,\omega\{H_i\}\}P(\max_{1\leq j\leq k} 2TC_j > 2z_{1-\alpha})}_{0} + P\{2,2,\omega\{H_i\}\}P(\max_{1\leq j\leq k} 2TC_j > 2z_{1-\alpha}) \\ =& \underbrace{P\{1,2,\omega\{H_i\}\}P(\max_{1\leq j\leq k} T_j > \sqrt{2z_{1-\alpha}})}_{0} + P\{2,2,\omega\{H_i\}\}P(\max_{1\leq j\leq k} T_j > \sqrt{2z_{1-\alpha}}) \\ =& P(\max_{1\leq j\leq k} T_j > \sqrt{2z_{1-\alpha}}) \end{aligned} \tag{5.12}$$

Finally we have

$$P(\max_{1\leq j\leq k} 2TC_j > 2z_{1-\alpha}) = P(\max_{1\leq j\leq k} T_j > \sqrt{2z_{1-\alpha}}) = \alpha \tag{5.13}$$

For the multivariate central weighted chi-square distribution with all degrees of freedom equal to one, we can use the correspond multivariate normal distribution with covariance matrix R to calculate the critical value asymptotically.

$$z_{1-\alpha} = 0.5\{\Phi_k^{-1}(p = 1 - \alpha; \mathbf{0}, \mathbf{R})\}^{\mathbf{2}} \tag{5.14}$$

which means

$$z_{1-\alpha} = 0.5Z^2_{k,1-\alpha} \tag{5.15}$$

$Z_{k,1-\alpha}$ is the α quantile for k-variate normal distribution.

5.2 Multiple Log-likelihood Test (MLT) with control of FWER

In this section, we will build a Multiple Log-likelihood Test (MLT) with control of FWER for Binomial order-restricted problems, by using the quantile which is developed in the last section. The test statistics are the differences discussed in the last section

$$TC_j = \log(L(\hat{g}(x|\hat{\theta}_j)), x \in H^j_A) - \log(L(\hat{g}(x|\hat{\theta}_0)), x \in H_0) \tag{5.16}$$

here, $\hat{\theta}_j = \{\hat{p_{j,0}}, ..., \hat{p_{j,k}}\}$ is the vector of the general estimators for elementary alternative H^j_A, $\hat{\theta}_0 = \{\hat{p}, ..., \hat{p}\}$ is the vector of the general estimators for null hypothesis H_0. Under the null hypothesis $TC = max\{TC_1, ..., TC_q\}$ is asymptotically q-variate weighted chi-square distributed, where q is the number of elementary alternatives. The choice of different likelihood estimators will affect the degrees of freedom of the weighted multivariate chi-square distribution. Here we discuss them in different situations. The chi-square distributions with SLE have one degree of freedom for all. We can simply prove it per definition.

5.2.1 Critical value

The critical value $z_{1-\alpha}$ is defined as

$$P\{[\max_{1\le j\le q} \log(L(\hat{g}(x), x \in H^j_A)) - \log(L(\hat{g}(x), x \in H_0)))] \ge z_{1-\alpha}\} = \alpha \tag{5.17}$$

Except the Epidemic-order restriction, *we use SLE to make the test statistics of MLT to be central q-variate chi-square distribution with all degree of freedom equal to one.* MLE is the only choice for the Epidemic-order restriction until now.

From last section, we have already known that the critical value of asymptotically q-variate chi-square distribution with all degree of freedom equal to one, can be calculated from a q-variate normal distribution.

$$z_{1-\alpha} = 0.5Z^2_{q,1-\alpha} \tag{5.18}$$

$Z_{q,1-\alpha}$ is the α quantile for q-variate normal distribution.

5.3 ORIC-lMLE, MHIC and MLT under order restriction

In this section, we will calculate the ORIC-lMLE, MHIC and MLT under different order restriction alternative models H^j_A.

5.3.1 Single Change-point order restriction

For ORIC-lMLE and MHIC methods, the GEP $\hat{\theta}_j$ are taken as the MLE $\tilde{\theta}_j$ which is equivalent to SLE $\bar{\theta}_j$ in this special case, i.e.

$$\hat{\theta}_j = \bar{\theta}_j = \tilde{\theta}_j = \{\tilde{p}_{j,1}, \tilde{p}_{j,2}, ..., \tilde{p}_{k,j}\} \tag{5.19}$$

with

$$\hat{p}_{j,i} = \tilde{p}_{j,i} = \min_{l \geq i} \max_{m \leq i} \frac{\sum_l^{v=m} \check{p}_{j,v}}{\sum_l^{v=m} n_v} \tag{5.20}$$

Under the null hypothesis $TC = max\{TC_1, ..., TC_q\}$, $q = k$ is asymptotically q-variate chi-square distributed. All of them have degree *one.*

The effect of using SLE and MLE are identical in this situation. We also have $MHIC = ORIC - lMLE$.

$$ORIC - lMLE = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\hat{p}_{j,i}^{x_i} * (1 - \hat{p}_{j,i})^{n_i - x_i}) - \sum_{m=0}^{r} w(r, m)(k - r + m) \tag{5.21}$$

since we have only one Change-point, $r = 1$ here implies

$$ORIC - lMLE = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\hat{p}_{j,i}^{x_i} * (1 - \hat{p}_{j,i})^{n_i - x_i}) - 1.5 \tag{5.22}$$

The MLT is given by

$$MLT = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\hat{p}_{j,i}^{x_i} * (1 - \hat{p}_{j,i})^{n_i - x_i}) - z_{1-\alpha} \tag{5.23}$$

Relationship of the ICs

Under alternatives H_A^j, we have

$$MLT - ORIC - lMLE \leq 1.5 - z_{1-\alpha} \tag{5.24}$$

5.3.2 Epidemic-order restriction

For AIC and NIC methods, $\hat{\theta}_j$ are taken as the MLE $\tilde{\theta}_j = \{\tilde{p}_{j,1}, \tilde{p}_{j,2}, ..., \tilde{p}_{k,j}\}$ where

$$\tilde{p}_{j,i} = \min_{l \geq i} \max_{m \leq i} \frac{\sum_l^{v=m} \check{p}_{j,v}}{\sum_l^{v=m} n_v} \tag{5.25}$$

Under the null hypothesis $TC = max\{TC_1, ..., TC_q\}$ with $q = (k-2)(k-1)$ is asymptotically central q-variate chi-square distributed. All of them have degree *two*.

Since we have two Change-points here

$$NIC = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\tilde{p}_{j,i}^{x_i} * (1 - \tilde{p}_{j,i})^{n_i - x_i}) - 2 - 2b(p'_1, p'_2) \tag{5.26}$$

here $b(p'_1, p'_2)$ is the extra penalty under Change-point order restriction for binomial data. p'_1 and p'_2 are the estimated proportion around the Change-points.

Relationship of the ICs

No MLT method is developed for this case. Under alternatives H_A, we have

$$NIC = AIC - 2b(p'_1, p'_2) \tag{5.27}$$

5.3.3 Simple-order restriction

For ORIC method, $\hat{\theta}_j$ are taken as the MLE $\tilde{\theta}_j$. Under the null hypothesis $TC = max\{TC_1, ..., TC_q\}$ with $q = k^2 - 1$ is asymptotically centered q-variate chi-square distributed with *different degree of freedom which depends on the elementary models.* We have $\hat{\theta}_j = \tilde{\theta}_j$ and

$$ORIC = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\tilde{p}_{j,i}^{x_i} * (1 - \tilde{p}_{j,i})^{n_i - x_i}) - \sum_{m=0}^{r} w(r, m)(k - r + m) \tag{5.28}$$

with

$$\tilde{p}_{j,i} = \min_{l \geq i} \max_{m \leq i} \frac{\sum_{l}^{v=m} \check{p}_{j,v}}{\sum_{l}^{v=m} n_v} \tag{5.29}$$

For MHIC method, $\hat{\theta}_j$ are taken as the SLE $\bar{\theta}_i$. Under the null hypothesis $TC = max\{TC_1, ..., TC_q\}$ with $q = k^2 - 1$ is asymptotically central q-variate chi-square

distributed. All of them have degree *one*.

$$MHIC = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\bar{p}_{j,i}^{x_i} * (1 - \bar{p}_{j,i})^{n_i - x_i}) - \sum_{m=0}^{r} w(r, m)(k - r + m) \tag{5.30}$$

with

$$\bar{p}_{j,i} = \hat{p} + \frac{c_{j,i}}{C_j} * \Delta \tag{5.31}$$

which is defined in Equation 4.44.

The MLT is given by using the same SLE as

$$MLT = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!} (\bar{p}_{j,i}^{x_i} * (1 - \bar{p}_{j,i})^{n_i - x_i}) - z_{1-\alpha} \tag{5.32}$$

As shown in the simulation study section, when the degree of freedom is higher than two, the effect of ORIC with MLE and MHIC with SLE are not identical. ORIC uses MLEs, so it is usually larger than MHIC

$$MHIC + 1.5 \preceq ORIC - lMLE + 1.5 \tag{5.33}$$

Relationship of the ICs

Under alternatives H_A^j, we have

$$MLT - ORIC - lMLE \leq 1.5 - z_{1-\alpha} \tag{5.34}$$

5.3.4 Simple-tree order restriction

For all methods, $\hat{\theta}_j$ are taken as the MLE $\tilde{\theta}_i$. Under the null hypothesis $TC = max\{TC_1, ..., TC_q\}$, $q = k$ is asymptotically q-variate chi-square distributed. All

of them have degree*one*. The effect of using SLE and MLE are identical in this situation. We also have $MHIC = ORIC - lMLE$. The MLE here is different as what we have before. For $i = 0$ we the MLE of $\hat{p}_{j,0} = \tilde{p}_{j,0}$ is

$$\hat{p}_{j,0} = \tilde{p}_{j,0} = \frac{x_0}{n_0} \tag{5.35}$$

For $i \neq 0$ the MLEs are

$$\hat{p}_{j,i} = \tilde{p}_{j,i} = max(\frac{x_i}{n_i}, \frac{x_0 + x_i}{n_0 + n_i}) \tag{5.36}$$

Here we use ORIC-lMLE too

$$ORIC - lMLE = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!}(\hat{p}_{j,i}^{x_i} * (1 - \hat{p}_{j,i})^{n_i - x_i}) - \sum_{m=0}^{r} w(r, m)(k - r + m) \tag{5.37}$$

since we have only inequity, $r = 1$ here implies

$$ORIC - lMLE = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!}(\hat{p}_{j,i}^{x_i} * (1 - \hat{p}_{j,i})^{n_i - x_i}) - 1.5 \tag{5.38}$$

The MLT is given by

$$MLT = \sum_{i=0}^{k} \log \frac{n_i!}{x_i!(n_i - x_i)!}(\hat{p}_{j,i}^{x_i} * (1 - \hat{p}_{j,i})^{n_i - x_i}) - z_{1-\alpha} \tag{5.39}$$

Relationship of the ICs

Under alternatives H_A^j, we have

$$MLT - ORIC - lMLE = 1.5 - z_{1-\alpha} \tag{5.40}$$

5.4 Relationship between MCT and MLT

In our last section, we have already seen that our MLT method has very similar behaviors as MCT when the normal approximation is true. In this section, the relationship between these two methods will be uncovered. Without order restriction, maximal likelihood estimator (MLE) is equivalent to least square estimator (LSE) for normal distributed data. When the data is not normal distributed, MLE has advantages over LSE.

Based on these MLE and SLE, MLT can be interpreted as information distance in general cases and when the data is normal distributed, MLT can be interpreted as quadratic distance (also true as information distance) while MCT can be interpreted as linear distance. According to Hirotsu (Hirotsu and Marumo, 2002): " ... the test statistic is also interpreted as the standardized maximum of the projections of the efficient score vector onto the k-1 corners of the polyhedral cone, where an efficient score vector is defined as the derivative of the log likelihood with respect to the parameter ... "

Observing the result from previous Single Change-point problems, we find that our MLT method is equivalent to MCT when the data is asymptotically normal distributed. This verifies Hirotsu's interpretation. The critical value of our new method is also calculated by taking quadratic of the critical value of correspond MCT.

But in principle, MCT and MLT are two different methods. MCT focused on test. The test statistics of MCT are multi-variate norm distributed and the critical value is easier to be calculated than MLT, which is designed to select the suitable model with control of FWER. In model selection procedure, Information Criterion is introduced to measure the distance between the estimation and the "truth". Under asymptotic normality assumption, Information distance can be interpreted as square distance and the distribution is multi-variate weighted chi-square distribution, which is quite complicated to calculate.

The square of the test statistics of MCT is multi-variate non-central chi-square distributed too. Under Single Change-point order restriction, both multi-variate

non-central chi-square distribution and multi-variate weighted chi-square distribution degenerate into multi-variate chi-square distribution with all degrees of freedom one. MCT and MLT are identical in this special case.

5.5 Algebraic space

In this section, former literatures of the Algebraic space will be reviewed. A description is given and an algebraic proof is developed for better understanding and it is also useful for the future study of the higher degrees of freedom case.

Algebraic space is a very useful tool to prove properties the test and model selection. Let the real true sample space to be an infinite dimensional real number space. The estimated models can be considered as a finite subspaces of the true space.

In the past, methods with consideration about orthogonal bases are developed (McDermott, 1999). This idea is simple and its computation is fast. But in many cases the orthogonal bases are hard to build. A more common corner vector space is introduced (Mukerjee et al., 1987). In the following part of this section, we show that MCT and MLT are identical under Single Change-point order restriction by using algebraic method. It would be better if the reader should have some basic knowledge of algebra. Or they can simply ignore this subsection, which is not important for understanding the whole method, and jump to next section.

Let random vector $X = \{X_0, X_1, ..., X_k\} \in \mathbb{R}^{k+1}$, $X_0, X_1, ..., X_k$ are normal distributed. X is multi-variate normal distributed with diagonal covariance matrix, $X \sim N(u, I\sigma^2)$. X can be considered as a random point in $(k+1)$ Euclidean space. Under certain order restriction, the null hypothesis $H_0 \in \mathbb{R}^{k+1}$ can be considered as a line and the alternative hypothesis can be considered as a closed convex cone, namely $H_A \in \mathbb{R}^{k+1}$, in this Euclidean space.

For example, as we have already discussed in former chapter, under Single Change-point order restriction, $H_0 = \{Y : y_0 = y_1 = ... = y_k\} \in \mathbb{R}^{k+1}$, $H_A^1 = \{Y : y_0 \leq y_1 = ... = y_k\} \in H_A$, ..., $H_A = \bigcup_{j=1}^{k} H_A^j \in \mathbb{R}^{k+1}$. If we also consider the contrasts as vec-

tors in $\mathbb{R}^{k+1}$, then each single contrast belongs to correspond elementary alternative model. For example, $C_1 = \{-k, 1, ..., 1\} \in H_A^1$.

Bretz(2005) gives a good interpretation of the relationship between the contrasts and the test statistics. Each elementary alternative H_A^j is a sub plane of H_A and vector $C_j \in H_A^j$ is contrast coefficients of the correspond elementary alternative. The contrast vector can be interpreted as prospected direction of the difference between vector X and its globe mean under null hypothesis. The maximal value of the MCT test statistics is bounded.

MCT defines the prediction as direction vector, while MLT defines the prediction in another way. Let $\hat{U_0}$ be the estimation of mean for X in H_0, $\hat{U_0} = \{\hat{u}, \hat{u}, ..., \hat{u}\} \in H_0, \hat{u} = \frac{1}{k+1}\sum_{i=0}^{k} X_i$, $(X - \hat{U_0}) \perp H_0$. By the same definition, we can defined the estimators under elementary alternatives. Let $\hat{U}_A^j$ be the mean estimator of X in H_A^j, $\hat{U}_A^j = \{\hat{u_0}, \hat{u_1}, ..., \hat{u_k}\} \in H_A^j$, $(X - \hat{U}_A^j) \perp H_A^j$. Further more, vector $(\hat{U}_A^j - \hat{U_0}) \in H_A^j$, so we have $(X - \hat{U}_A^j) \perp (\hat{U}_A^j - \hat{U_0})$. In $\mathbb{R}^{k+1}$, points $X, \hat{U}_A^j, \hat{U_0}$ form a right triangle. Define the Norm in $\mathbb{R}^{k+1}$ as $\|Z\| = \sqrt{\sum_0^{k+1} z_i^2}$. We have,

$$\|(X - \hat{U}_A^j)\|^2 + \|(\hat{U}_A^j - \hat{U_0})\|^2 = \|(X - \hat{U_0})\|^2 \tag{5.41}$$

The Log-likelihood $L(H_A^j)$can be interpreted as half quadratic distance between the observation X and estimation $\hat{X}_A^j$.

$$2L(H_A^j) = -\sum_{i=0}^{k} \frac{(X_i - \hat{u_i})^2}{\sigma^2/n} = -\frac{n}{\sigma^2}\|X - \hat{U}_A^j\|^2 \tag{5.42}$$

$$2L(H_0) = -\sum_{i=0}^{k} \frac{(X_i - \hat{u})^2}{\sigma^2/n} = -\frac{n}{\sigma^2}\|X - \hat{U_0}\|^2 \tag{5.43}$$

Until now we have two sides of the right angle. The third one is given as following: Given two vectors $A = \{a_0, ...a_k\} \in \mathbb{R}^{k+1}$ and $B = \{b_0, ...b_k\} \in \mathbb{R}^{k+1}$, we can define

the inner product $\langle A, B\rangle = \sum_{i=0}^{k}(a_i b_i)$ and $\langle A, A\rangle = \|A\|^2$. Then $\sum_{i=0}^{k}(c_i x_i)$ can be noted as $\langle C, X\rangle$. Because of $\langle C, \hat{X}\rangle = 0$, we get the following,

$$\langle C, X\rangle = \langle C, X\rangle - 0 = \langle C, X\rangle - \langle C, \hat{U}_0\rangle = \langle C, (X - \hat{U}_0)\rangle \tag{5.44}$$

since $C_j, \hat{U}_A^j \in H_A^j$, $(\hat{U}_A^j - \hat{U}_0) \in H_A^j$, $(X - \hat{U}_A^j) \perp (\hat{U}_A^j - \hat{U}_0)$ we have,

$$0 \leq \angle(\hat{U}_A^j - \hat{U}_0, X - \hat{U}_0) \leq \angle(C_j, (X - \hat{U}_0)) \leq \pi/2 \tag{5.45}$$

The best prediction of vector $X - \hat{U}_0$ in plane is $\hat{U}_A^j - \hat{U}_0$. The equality of our last equation holds when,

$$C_j = \lambda \frac{\hat{U}_A^j - \hat{U}_0}{\|\hat{U}_A^j - \hat{U}_0\|} \tag{5.46}$$

Here, λ is any positive constant.

Under the condition of known variance and equal sample size, contrast test statistics for alternative H_A^j can be written as,

$$\begin{aligned} MCT(H_A^j) &= \frac{\sum \frac{c_i}{n} x_i}{\sqrt{\{\sigma^2 \sum \frac{c_i^2}{n}\}}} = \sqrt{\frac{n}{\sigma^2}} * \frac{\langle C_j, (X - \hat{U}_0)\rangle}{\sqrt{\langle C_j, C_j\rangle}} \\ &= \sqrt{\frac{n}{\sigma^2}} \frac{\langle C_j, (X - \hat{U}_0)\rangle}{\|C_j, C_j\| \|X - \hat{U}_0\|} \|X - \hat{U}_0\| \\ &= \sqrt{\frac{n}{\sigma^2}} \cos(C_j, (X - \hat{U}_0)) * \|X - \hat{U}_0\| \\ &\leq \sqrt{\frac{n}{\sigma^2}} \cos(\angle(\hat{U}_A^j - \hat{U}_0, X - \hat{U}_0)) * \|X - \hat{U}_0\| \\ &= \sqrt{\frac{n}{\sigma^2}} \|\hat{U}_A^j - \hat{U}_0\| \end{aligned} \tag{5.47}$$

Putting Equation 5.42, 5.43 and 5.47 into Equation 5.41, finally we have,

$$\begin{aligned} \|(X-\hat{U}_A^j)\|^2 + \|(\hat{U}_A^j-\hat{U}_0)\|^2 &= \|(X-\hat{U}_0)\|^2 \\ \frac{n}{\sigma^2}\|(X-\hat{U}_A^j)\|^2 + \frac{n}{\sigma^2}\|(\hat{U}_A^j-\hat{U}_0)\|^2 &= \frac{n}{\sigma^2}\|(X-\hat{U}_0)\|^2 \\ 2*(L(H_A^j)-L(H_0)) &\geq MCT(H_A^j)^2 \end{aligned} \tag{5.48}$$

Under Single Change-point order restriction, C_j satisfies the condition in Equation 5.46 and the equality holds

$$2*(L(H_A^j)-L(H_0)) = MCT(H_A^j)^2 \tag{5.49}$$

This verifies the same result from Equation 5.10. This relationship holds true, after adjusted for unequal sample size case.

According to Robertson et al.(1988): "...it can be shown that the LRT statistic may be expressed as the maximum of an infinite number of contrast statistics." Our result shows that the LRT statistics may also be considered as the maxinf of correspond local log-likelihood statistics.

An virtual example is given as following (here we enlarge the sample size to get the asymptotic normality),

	X_0	X_1	X_2
Present x_i	200	200	300
Absent $n_i - x_i$	300	300	200
Total n_i	500	500	500
$\hat{p}_i$	0.40	0.40	0.60

	H_0	H_A^1	H_A^2
MCT (H_A^j)		3.65	7.31
MCT $(H_A^j)^2$		13.39286	53.57143
L(H)	-36.807100	-28.575150	-8.439637
2*L(H)	-73.61420	-57.15030	-16.87927
2*($L(H_A^i)$-L(H_0))		16.4	56.7

Table 5.1: Example of the relationship when $k = 3$.

The hypotheses of the null and alternatives are,

$$\begin{aligned} H_0 &: p_0 = p_1 = p_2 \\ H_A^1 &: p_0 < p_1 = p_2 \\ H_A^2 &: p_0 = p_1 < p_2 \end{aligned} \tag{5.50}$$

The calculated test statistics are in Table 5.1. The relationship for this 3 dimension case is also described in Figure 5.1 and 5.2. In Figure 5.1, we give an overview of the three models. The axes are labeled by the observed proportions x_0, x_1, x_2. The red plane is elementary alternative $H_A^1 = \{Y : y_0 \leq y_1 = y_2\}$ and the green plane is elementary alternative $H_A^2 = \{Y : y_0 = y_1 \leq= y_2\}$. The null model $H_0 = \{Y : y_0 = y_1 = y_2\}$ is a line which lies in between the other two models i.e. H_0 is the boundary of these two open sets. The three balls are confidence region for these three model.

The enlarged picture is shown in Figure 5.2. In this picture X is the observed data vector. $H0 \in H_0$, $H1 \in H_A^1$ and $H2 \in H_A^2$ are the predicted data vector on null hypothesis space H_0, elementary alternative plane H_A^1 and H_A^2, i.e.

$$(X - H0) \perp H_0, (X - H1) \perp H_A^1 \text{ and } (X - H_A^2) \perp H_A^2 \tag{5.51}$$

The blue, cyan and purple ellipse are simulated confidence regions from the predicted data vector $H0$, $H1$ and $H2$. In this picture, we have two triangles which can be

noted as $(X, H2, H0)$ and $(X, H1, H0)$. It is simply to prove that

$$(X - H2) \perp (H2 - H0) \text{ and } (X - H1) \perp (H1 - H0) \tag{5.52}$$

By the triangle rules we have

$$\|(X - H0)\|^2 - \|(X - H2)\|^2 = \|(H2 - H0)\|^2 \tag{5.53}$$

and

$$\|(X - H0)\|^2 - \|(X - H1)\|^2 = \|(H1 - H0)\|^2 \tag{5.54}$$

Putting Equation 5.42, 5.43 and 5.47 into these equation, finally we have

$$\begin{aligned} 2 * (L(H_A^1) - L(H_0)) &= MCT(H_A^1)^2 \\ 2 * (L(H_A^2) - L(H_0)) &= MCT(H_A^2)^2 \end{aligned} \tag{5.55}$$

and check the real value calculated from Table 5.1, we have

$$\begin{aligned} 2 * (L(H_A^1) - L(H_0)) &= 16.4, MCT(H_A^1)^2 = 13.4 \\ 2 * (L(H_A^2) - L(H_0)) &= 56.7, MCT(H_A^2)^2 = 53.6 \end{aligned} \tag{5.56}$$

These results verify Equation 5.56.

5.6 Model selection with control of FWER for MLT

When the null hypotheses is rejected, we select the model which has the smallest KL distance as the best model. MLT is a simultaneous procedure which can do model selection after the test by meaning of KL distance. Furthermore, for given elementary

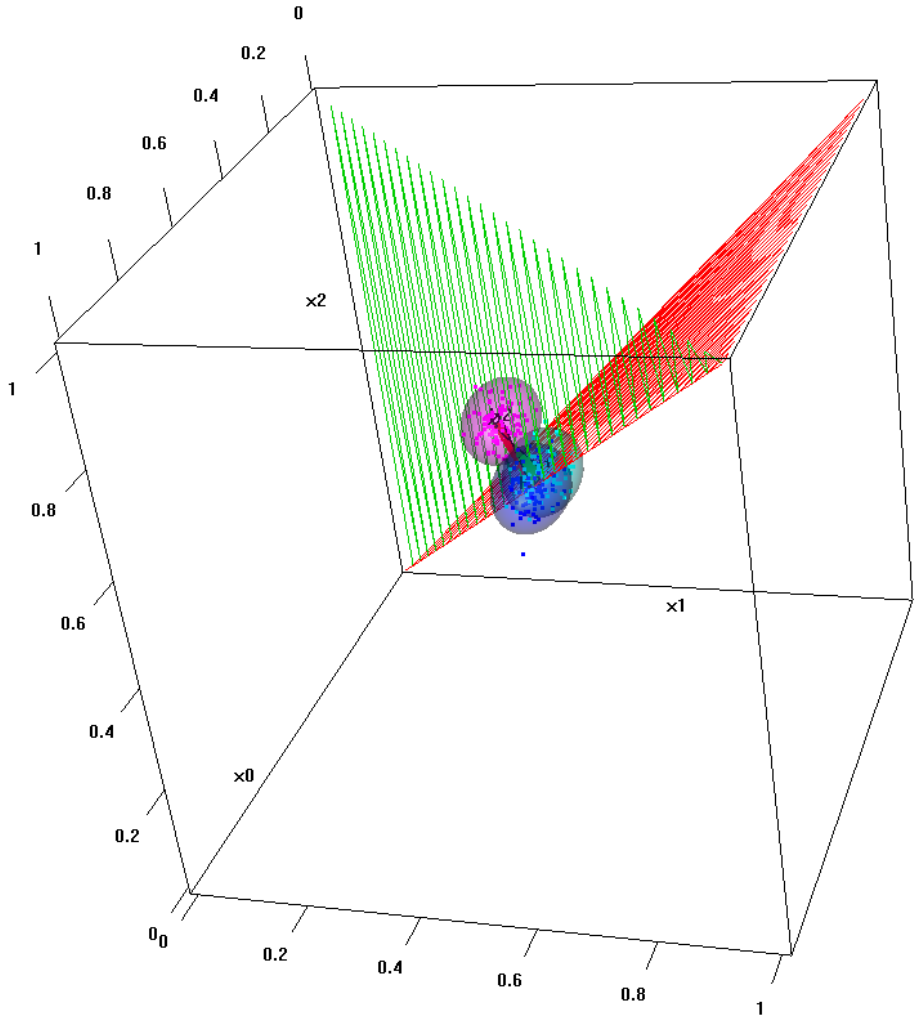

Figure 5.1: Three dimension plot for simulated binomial data k=2

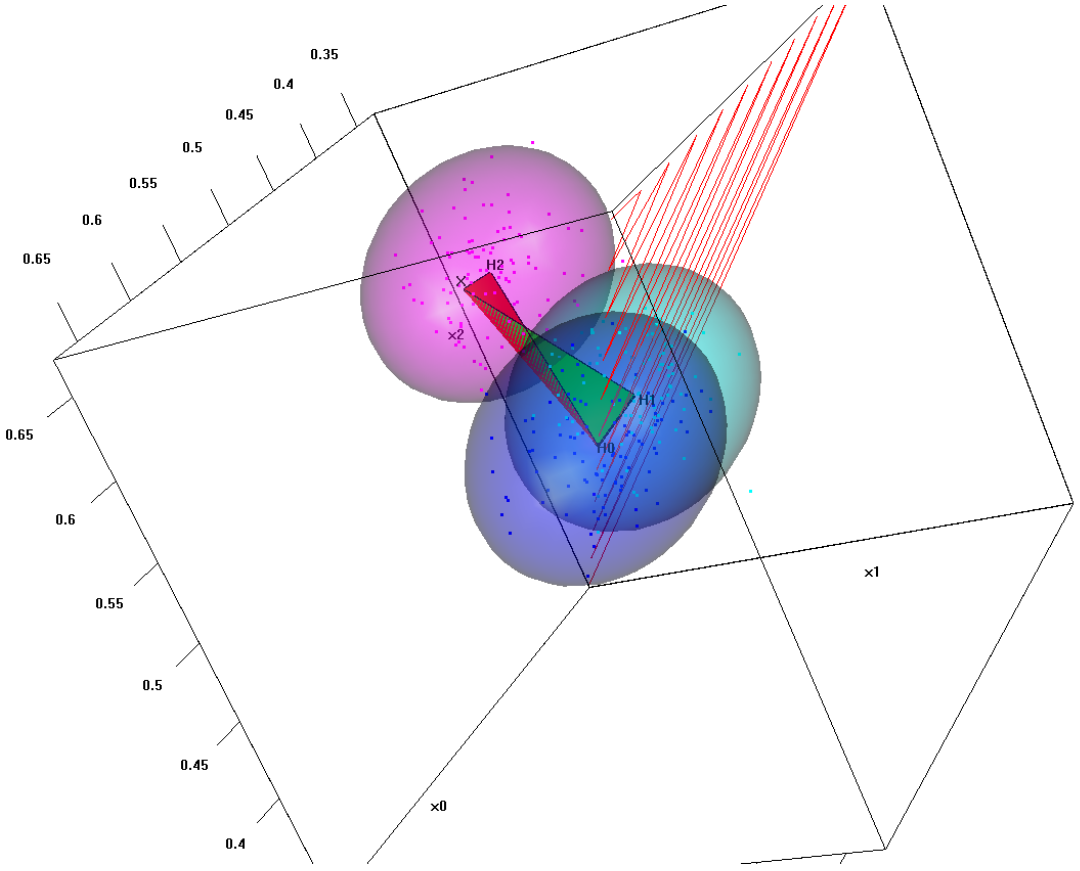

Figure 5.2: Enlarged three dimension plot for simulated binomial data k=2

.

alternative model the ICs of MLT and MHIC have the following relationship

$$MLT - MHIC = 1.5 - z_{1-\alpha} \leq ORIC - lMLE - MHIC \tag{5.57}$$

For given null model, the ICs of AIC, ORIC, ORIC-lMLE, MLT and MHIC have the following relationship

$$MLT = MHIC = AIC = ORIC = ORIC - lMLE \tag{5.58}$$

Chapter 6

Power study and simulation

As we have seen in the last chapter, there is a connection between MCT and MLT that $T_{max}^2 \leq 2TC$. In this chapter, we will use this relationship to calculate the power, the correct model selection rate and the misclassification rate of MLT. A simulation study will also be given to compare the model selection methods under different order restrictions.

6.1 Expressions

The power is the rate that the test successfully rejects the null hypothesis with certain error level, when the alternative model is true. Under order restriction, no uniformly powerful test exists Bretz (1999). Most of the tests are specialized by their background assumption, e.g. MLT for Single Change-point are very "powerful", if there is one and only one Change-point in the uncovered data structure. In this case, we see that MLT achieves its "maximum power". In certain other contradict cases, MLT will not behave as the way we expected. One of these cases, e.g. Simple-order and Simple-tree order let these MLT achieve its "minimum power". At the same time, ORIC is more sensitive to detect the Simple-order. We will see the evidence in the simulation section.

The correct model selection rate (CR) is the rate that the test successfully selects the

right model. The CR of test method is usually lower than those of model selection method, because of the error control.

The Misclassification rate(MR) describes how often that we select the wrong model. It can be easily proved that the sum of MR and CR is equal to one.

6.1.1 Expression of the power

Similarly as MCT, the asymptotic power of MLT can be calculated by,

$$\begin{aligned} & P(\max_{1\leqslant l\leqslant q}\{TC_l \geqslant z_{1-\alpha}\} \mid H_A) \\ =& 1 - P(TC_1 \leq z_{1-\alpha}, \text{and...and, } TC_q \leq z_{1-\alpha} \mid H_A) \end{aligned} \tag{6.1}$$

under alternatives it is multivariate non-central chi-square distributed. But for Single Change-point problem and any other problem that uses SLE as estimators, the power of them can be calculated by non-central normal distribution

$$\begin{aligned} =& 1 - P(T_1 \leq \sqrt{2z_{1-\alpha}}, \text{and...and, } T_q \leq \sqrt{2z_{1-\alpha}} \mid H_A) \\ =& 1 - \Phi_q((\sqrt{2z_{1-\alpha}})\text{diag}(\frac{1}{v_1}, ..., \frac{1}{v_q}); \mathbf{e}, \mathbf{R}) \\ =& 1 - \Phi_q((z_{q,1-\alpha})\text{diag}(\frac{1}{v_1}, ..., \frac{1}{v_q}); \mathbf{e}, \mathbf{R}) \end{aligned} \tag{6.2}$$

Here, we use the relationship of $z_{q,1-\alpha} = \sqrt{2z_{1-\alpha}}$ in Equation 5.15. Under the condition of Single Change-point and large enough sample size, our new method and MCT share the asymptotic power. Furthermore, if we replace $z_{1-\alpha}$ by the penalty term (i.e. $c_{ORIC-lMLE} = 1.5$, $c_{AIC} = 2$) of ORIC-lMLE or AIC methods, we can calculate the equivalent power of these methods. Examples will be given in the simulation study and summary sections.

6.1.2 Correct model selection rate (CR)

Correct model selection rate(CR) describes how often that we select the correct model. We could use mathematical language to define it as "CR=P(we select model i| the true model is i)".

The CR of the null model has already been well described by FWER. So here we focus on the CR for the alternative models, which is also called sensitivity. It is the model selection rate of H_A^j, when the real model is H_A^j. Two requirements must be fulfilled: First, the globe alternative should be rejected. Second, the test statistics should be the highest among them. As presented in the previous section, our MLT shares the likelihood estimator and critical value of MCT. Furthermore, they shares the power expression under Single Change-point order restriction. In the next part of this section, we will show how to use the CR of MCT to approximate the CR of MLT.

The CR of MCT is,

$$\begin{aligned} &P(\text{select}H_A^j|H_A^j) \\ &= P(T_j > T_1, ..., T_j > T_{j-1}, T_j > z_{q,1-\alpha}, T_j > T_{j+1}, ..., T_j > T_q|H_A^j) \end{aligned} \tag{6.3}$$

This is a multi-variated normal probability with random upper limits. It is not easy to calculate, so we make the following transformation. Let

$$T_l' = \begin{cases} T_l - T_j, & 0 \leq l \leq j-1 \\ z_{q,1-\alpha} - T_j, & l = j \\ T_l - T_j, & j+1 \leq l \leq q \end{cases} \tag{6.4}$$

here $e' = (E(T_1'), ..., E(T_q'))$ and $v' = (V(T_1'), ..., V(T_q'))$ are the means and variances of new vector $T' = \{T_1', T_2', ..., T_q'\}$ under H_A^j. R' is the new correlation matrix. We use simulated value of the variance v' instead of calculating it. The CR of MCT is

transformed to

$$\begin{aligned} & P(\text{select}H_A^j|H_A^j) \\ =& P(T_1' < 0, ..., T_{j-1}' < 0, T_j' < 0, T_{j+1}' < 0, ..., T_q' < 0|H_A^j) \\ =& \Phi_q((\mathbf{e}')\text{diag}(\frac{1}{v_1'}, ..., \frac{1}{v_q'}); \mathbf{0}, \mathbf{R}') \end{aligned} \tag{6.5}$$

The CR of MLT is non-centred chi-square distributed, but we can also use the CR of MCT as an approximation.

$$\begin{aligned} & P(\text{select}H_A^j|H_A^j) \\ & = P(TC_j > TC_1, ..., TC_j > TC_{j-1}, TC_j > z_{1-\alpha}, TC_j > TC_{j+1}, ..., TC_j > TC_q|H_A^j) \\ & \approx P(T_j > T_1, ..., T_j > T_{j-1}, T_j > \sqrt{2z_{1-\alpha}}, T_j > T_{j+1}, ..., T_j > T_q|H_A^j) \end{aligned} \tag{6.6}$$

Similarly as we get the power for ORIC-lMLE and AIC methods, the equivalent CR of them can be calculated by replacing $z_{q,1-\alpha}$ with $\sqrt{2c_{AIC}}$ or $\sqrt{2c_{ORIC-lMLE}}$ in the CR equation of MCT.

6.1.3 Misclassification rate(MR)

The Misclassification rate(MR) describes how often that we select the wrong model. We could use mathematical language to define it as

$$P(\text{select}H_A^j|\bar{H}_A^j) \tag{6.7}$$

here, $\bar{H}_A^j$ is the complementary set of event that "The true model is H_A^j", i.e. $\bar{H}_A^j = (\bigcup_{i \neq j} H_A^i) \bigcup H_0$ The MR for MCT can be described as

$$\begin{aligned} & P(\text{select}H_A^j|\bar{H}_A^j) \\ =& P(T_j > T_1, ..., T_j > T_{j-1}, T_j > z_{q,1-\alpha}, T_j > T_{j+1}, ..., T_j > T_q|\bar{H}_A^j) \end{aligned} \tag{6.8}$$

No algorithm available until now to calculate this probability. We could only calculate the upper bound of this probability as

$$P(\text{select}H_A^j|\bar{H}_A^j) \leq P(H_A^j|\bar{H}_0) + \sum_{i \neq j} P(H_A^j|\bar{H}_A^i) \tag{6.9}$$

The equality holds if and only if all the models are not overlapped. Furthermore, we cannot use the MR of MCT to approximate MR of MLT. If the value of MCT is small, the error will be relatively large. So the problem of calculating MR for MLT is left as the second open question in this thesis.

6.2 Simulation study

In this chapter, we will make a simulation study for the four types of order restriction we discussed before. MCT, MLT, ORIC-lMLE and MHIC methods will be compared here. For Single Change-point order restriction we use ORIC-lMLE which is an improvement of ORIC.

6.2.1 Single Change-point order

We generate 10000 random binomial data for $k = 2$ with means $p_0 = 0.4$, $p_1 = 0.4$, $p_2 = 0.4 + \Delta = 0.6$, and simple size 25, 50, 100. There is only one Change-point. Therefore MHIC and ORIC-lMLE methods are equivalent for this situation. The model selection rate of MCT, ORIC-lMLE and MLT are compared in Picture 6.1. In these 3 pictures we see that the correct model selection rate (CR) increases when the sample size or the value of Δ increases. It can be sure that the CR of these three methods will convergent to one if the sample size or Δ is large enough. In the third one of Picture 6.1, we see that MLT and MCT are almost identical. This verifies our former proof that MLT and MCT are asymptotical equivalent under Single Change-point order. However, MLT has a little bit higher CR than MCT

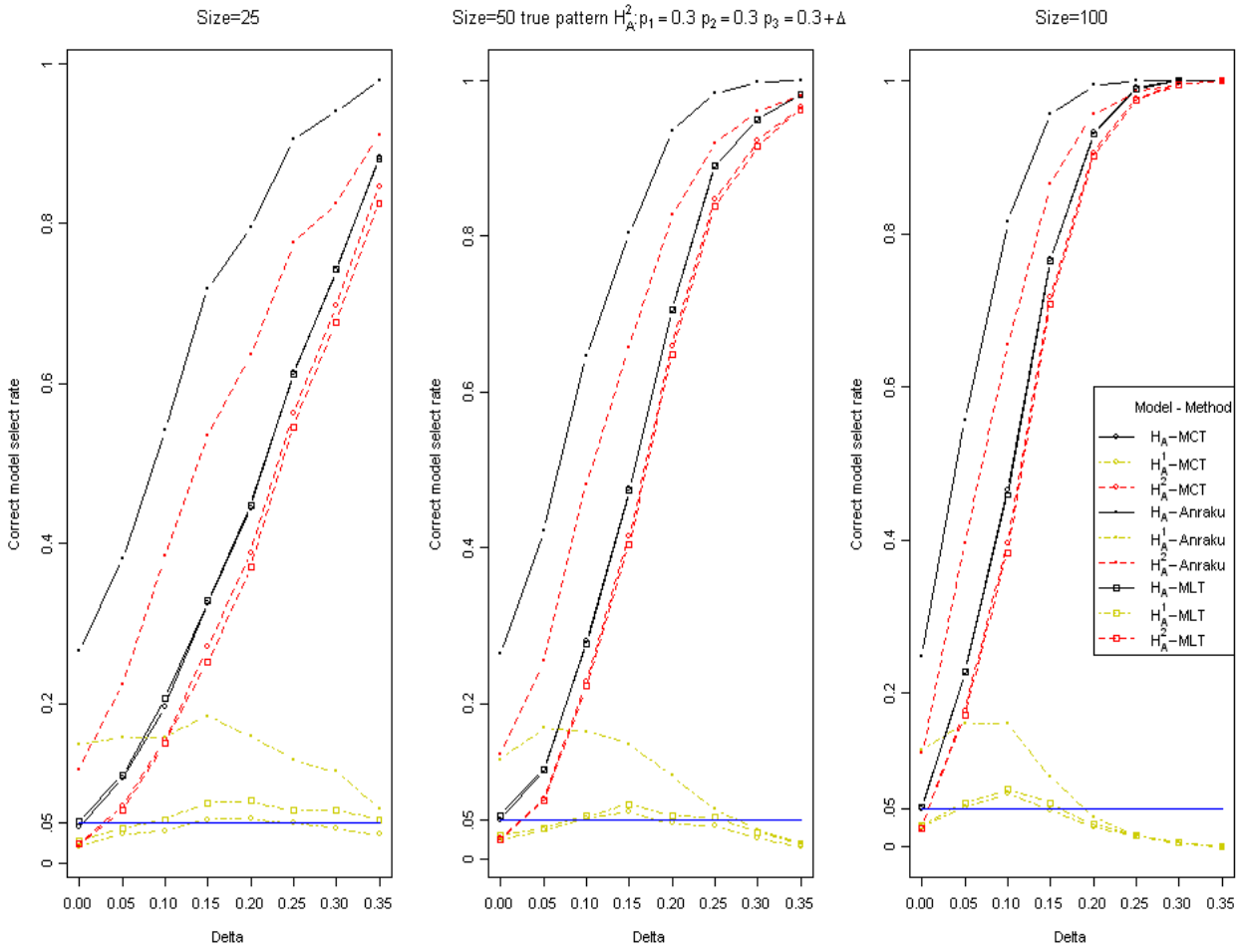

Figure 6.1: Simulation of power and model selection rate

when the sample size is smaller (see in the first one of Picture 6.1). This verifies our former conclusion that KL distance is more useful in model selection than general difference or LSE. Both of MLT and MCT control FWER in these three different situations, while ORIC-lMLE cannot. However, ORIC-lMLE has a higher CR than the other two.

When non-centrality is large enough, both estimators of MCT and ORIC-lMLE have similar variance. But in general, the estimator chosen by ORIC-lMLE has smaller variance than MCT (Chaudhuri and Perlman, 2005). So ORIC-lMLE method has larger identification rate than MCT in identifying the correct alternative models.

For rejecting the null hypothesis, the situation is quite different. ORIC-lMLE

method is a sensitive one to detect the Change-point, but it is too sensitive that it has very high false positive error rate α . It selects model without α control. A false positive is reported. This over-estimation is kind of "underperformance", which is discussed by Roberts and Martin (2006) and Zucchini (2000). Over-estimation under the null is a common problem of IC methods, because they are not designed to make a test. In our example here, ORIC-lMLE identifies only around 60% of the null model when $\Delta = 0$, while MCT and our method identifies 95% of the null model. The following tables give further examples of the correct model selection rate over different alternatives.

For asymptotic normal case, we generate 10000 random binomial data for $k = 5$ with means $p_0 = ... = p_{j-1} = 0.4$, $p_j = ... = p_k = 0.4 + \Delta = 0.6$, and the sample size is 50. The result is shown in Table 6.1. As shown in the table, the correct model selection rates vary from different patterns. This result also verifies the theoretical value of the power and correct model selection rate. MHIC is not considered here, because it is totally identical with ORIC-lMLE under Single Change-point order restriction.

For anastomotic normal case, we generate 10000 random binomial data for $k = 5$ with means $p_0 = ... = p_{j-1} = 0.01$, $p_j = ... = p_k = 0.01 + \Delta = 0.07$, and the sample size is 100. The result is shown in Table 6.2. From this simulation example, we can find out that MLT is stable in selecting model H_A^1 and H_A^2 when the sample size is getting smaller. In the situation when normality is not fulfilled, MLT still controls the FWER, and also has good CR, while MCT has problems in model selection.

In Table 6.3, we study the equivalent power of ORIC-lMLE. Let the critical value equal to the penalty term of ORIC-lMLE, i.e. $z_{1-\alpha} = 1.5$. Using the relationship between MCT and MLT $Z_{q,1-\alpha} = \sqrt{2z_{1-\alpha}}$. By solving $Z_{6,1-\alpha} = \sqrt{2 * 1.5}$, we get $\alpha = 0.41$. Now, we change the α rate of MCT into 41%, namely "MCT059", and other situations are totally the same as the first simulation. The "MCT059" has similar behaviors as MLT method.

In the last table (Table 6.4), we also try the unbalanced sample size. The power estimation of model H_A^5 has problems. Under a small sample size, the assumption

Alternatives	Meth.	$Asy.H_0$	H_0	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5	Sel.rate
.4/.4/.4/.4/.4/.4	MLT	.9498	.9479	.0110	.0114	.0098	.0101	.0098	.9498
.4/.4/.4/.4/.4/.4	MCT	.9498	.9486	.0111	.0122	.0096	.0099	.0086	.9498
.4/.4/.4/.4/.4/.4	ORIC-lMLE	.5916	.5933	.0927	.0711	.0674	.0776	.0979	.5916
.6/.6/.6/.6/.6/.6	MLT	.9500	.9465	.0110	.0098	.0114	.0102	.0111	.9500
.6/.6/.6/.6/.6/.6	MCT	.9500	.9483	.0089	.0090	.0116	.0109	.0113	.9500
.6/.6/.6/.6/.6/.6	ORIC-lMLE	.5916	.5869	.0995	.0793	.0703	.0692	.0948	.5916
.4/.6/.6/.6/.6/.6	MLT	.2943	.3037	.5857	.0586	.0264	.0141	.0115	.6249
.4/.6/.6/.6/.6/.6	MCT	.2943	.3031	.5902	.0586	.0260	.0125	.0096	.6249
.4/.6/.6/.6/.6/.6	ORIC-lMLE	.0285	.0288	.7788	.0906	.0450	.0295	.0273	.8158
.4/.4/.6/.6/.6/.6	MLT	.1036	.1206	.0541	.7229	.0707	.0217	.0100	.7304
.4/.4/.6/.6/.6/.6	MCT	.1036	.1203	.0531	.7264	.0703	.0209	.0090	.7304
.4/.4/.6/.6/.6/.6	ORIC-lMLE	.0044	.0054	.0676	.7981	.0835	.0297	.0157	.7976
.4/.4/.4/.6/.6/.6	MLT	.0708	.0800	.0183	.0683	.7474	.0675	.0185	.7643
.4/.4/.4/.6/.6/.6	MCT	.0708	.0802	.0165	.0676	.7510	.0668	.0179	.7643
.4/.4/.4/.6/.6/.6	ORIC-lMLE	.0024	.0026	.0229	.0770	.7971	.0771	.0233	.8188
.4/.4/.4/.4/.6/.6	MLT	.1036	.1148	.0106	.0252	.0672	.7238	.0584	.7182
.4/.4/.4/.4/.6/.6	MCT	.1036	.1139	.0095	.0228	.0675	.7285	.0578	.7182
.4/.4/.4/.4/.6/.6	ORIC-lMLE	.0044	.0052	.0163	.0318	.0807	.7971	.0689	.8023
.4/.4/.4/.4/.4/.6	MLT	.2944	.3104	.0131	.0140	.0240	.0595	.5790	.6240
.4/.4/.4/.4/.4/.6	MCT	.2944	.3099	.0119	.0127	.0230	.0614	.5811	.6240
.4/.4/.4/.4/.4/.6	ORIC-lMLE	.0285	.0301	.0280	.0301	.0449	.0920	.7749	.8077

Table 6.1: 10000 random binomial data for $k = 5$, proportions $p_0 = ... = p_{j-1} = 0.4$, $p_j = ... = p_k = 0.6$, sample size n_i is 50.

Alternatives	Meth.	j	$Asy.H_0$	H_0	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5
.01/.01/.01/.01/.01/.01	MLT	-	0.9498	0.9447	0.0145	0.0098	0.0079	0.0108	0.0123
.01/.01/.01/.01/.01/.01	MCT	-	0.9498	0.9498	0.0054	0.0064	0.0081	0.0122	0.0181
.01/.01/.01/.01/.01/.01	ORIC-lMLE	-	0.5916	0.5806	0.1110	0.0782	0.0635	0.0714	0.0953
.07/.07/.07/.07/.07/.07	MLT	-	0.9500	0.9239	0.0002	0.0220	0.0192	0.0170	0.0177
.07/.07/.07/.07/.07/.07	MCT	-	0.9500	0.9351	0.0000	0.0007	0.0087	0.0194	0.0361
.07/.07/.07/.07/.07/.07	ORIC-lMLE	-	0.5915	0.5066	0.1821	0.0936	0.0834	0.0602	0.0741
.01/.07/.07/.07/.07/.07	MLT	1	0.0557	**0.1871**	0.7118	0.0582	0.0224	0.0120	0.0085
.01/.07/.07/.07/.07/.07	MCT	1	0.0557	**0.3256**	**0.4860**	0.0894	0.0450	0.0291	0.0249
.01/.07/.07/.07/.07/.07	ORIC-lMLE	1	0.0017	0.0062	0.8575	0.0752	0.0293	0.0180	0.0138
.01/.01/.07/.07/.07/.07	MLT	2	0.0047	0.0396	0.0386	0.8455	0.0556	0.0148	0.0059
.01/.01/.07/.07/.07/.07	MCT	2	0.0047	0.0575	0.0044	**0.7677**	0.1017	0.0440	0.0247
.01/.01/.07/.07/.07/.07	ORIC-lMLE	2	0.0000	0.0005	0.0410	0.8757	0.0592	0.0161	0.0075
.01/.01/.01/.07/.07/.07	MLT	3	0.0019	0.0187	0.0088	0.0515	0.8586	0.0505	0.0119
.01/.01/.01/.07/.07/.07	MCT	3	0.0019	0.0238	0.0002	0.0121	0.8211	0.1042	0.0386
.01/.01/.01/.07/.07/.07	ORIC-lMLE	3	0.0000	0.0005	0.0096	0.0531	0.8714	0.0523	0.0131
.01/.01/.01/.01/.07/.07	MLT	4	0.0047	0.0358	0.0035	0.0176	0.0516	0.8461	0.0454
.01/.01/.01/.01/.07/.07	MCT	4	0.0047	0.0319	0.0000	0.0021	0.0170	0.8457	0.1033
.01/.01/.01/.01/.07/.07	ORIC-lMLE	4	0.0000	0.0017	0.0073	0.0187	0.0537	0.8686	0.0500
.01/.01/.01/.01/.01/.07	MLT	5	0.0556	0.1366	0.0026	0.0165	0.0227	0.0529	0.7687
.01/.01/.01/.01/.01/.07	MCT	5	0.0556	0.1041	0.0000	0.0010	0.0056	0.0245	0.8648
.01/.01/.01/.01/.01/.07	ORIC-lMLE	5	0.0017	0.0143	0.0174	0.0207	0.0303	0.0653	0.8520

Table 6.2: 10000 random binomial data for $k = 5$, proportions $p_0 = ... = p_{j-1} = 0.01$,$p_j = ... = p_k = 0.07$, sample size n_i is 100.

Alternatives	Meth.	$Asy.H_0$	H_0	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5
0.6/0.6/0.6/0.6/0.6/0.6	MLT	0.9501	0.9472	0.0113	0.0098	0.0108	0.0091	0.0118
0.6/0.6/0.6/0.6/0.6/0.6	MCT059	0.5901	0.5846	0.1062	0.0710	0.0719	0.0760	0.0903
0.6/0.6/0.6/0.6/0.6/0.6	ORIC-lMLE	0.5916	0.5895	0.1019	0.0677	0.0718	0.0762	0.0929
0.4/0.4/0.4/0.4/0.6/0.6	MLT	0.1034	0.1107	0.0130	0.0243	0.0701	0.7246	0.0573
0.4/0.4/0.4/0.4/0.6/0.6	MCT059	0.0044	0.0053	0.0147	0.0279	0.0808	0.8039	0.0674
0.4/0.4/0.4/0.4/0.6/0.6	ORIC-lMLE	0.0044	0.0053	0.0161	0.0308	0.0808	0.7987	0.0683
0.4/0.4/0.4/0.4/0.4/0.6	MLT	0.2941	0.2972	0.0123	0.0156	0.0233	0.0638	0.5878
0.4/0.4/0.4/0.4/0.4/0.6	MCT059	0.0282	0.0265	0.0279	0.0307	0.0386	0.0929	0.7834
0.4/0.4/0.4/0.4/0.4/0.6	ORIC-lMLE	0.0285	0.0273	0.0292	0.0311	0.0400	0.0912	0.7812

Table 6.3: ORIC-lMLE is equivalent to a MCT with lower control of FWER (=0.41) under the given situation.

Alternatives	Methods	$Asy.H_0$	H_0	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5
0.4/0.4/0.4/0.6/0.6/0.6	MLT	-	0.0927	0.0250	0.0669	0.7343	0.0633	0.0178
0.4/0.4/0.4/0.6/0.6/0.6	MCT059	-	0.0072	0.0905	0.1340	0.7258	0.0256	0.0169
0.4/0.4/0.4/0.6/0.6/0.6	ORIC-lMLE	-	0.0066	0.0315	0.0779	0.7894	0.0711	0.0235
0.4/0.4/0.4/0.4/0.4/0.6	MLT	0.2946	0.5155	0.0178	0.0197	0.0294	0.0760	0.3416
0.4/0.4/0.4/0.4/0.4/0.6	MCT059	-	0.0881	0.0702	0.0637	0.0755	0.1212	0.5813
0.4/0.4/0.4/0.4/0.4/0.6	ORIC-lMLE	-	0.1171	0.0472	0.0445	0.0608	0.1324	0.5980

Table 6.4: 10000 random data with unbanlanced sample size 100/50/50/50/25/25

of asymptotic normal is not fulfilled.

6.2.2 Epidemic-order

One application of model selection under Epidemic-order restriction is DNA-motif finding, which requires fast and efficient algorithm to find out the position of Change-points. Both MCT and IC methods fulfill this requirement. However, IC method works faster in motif finding because it simplifies the critical value calculation with a constant penalty term. Here we mainly present the result of NIC and MLT. But MCT still have the advantage that it can calculate the asymptotic power. For the power study of log-likelihood test, three simulations are generated 10,000 times for each.

As shown in Figure 6.2, the data under the null is generated. All the positions have the same parameter, noted as p_1. In this situation, at 99% of the time, the null hypotheses are accepted. When the sequence length becomes larger, the variation

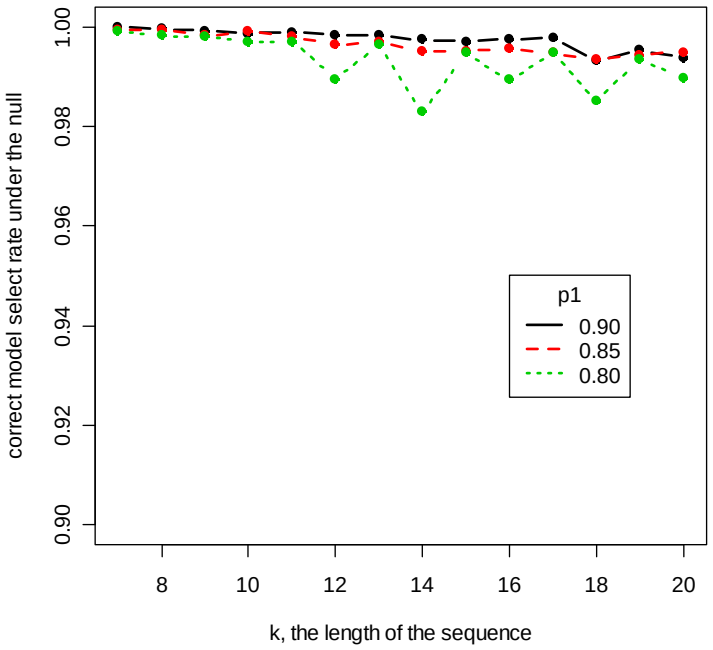

Figure 6.2: According to the null, the finding rate of NIC is acceptable

of the data becomes greater and the finding rate becomes smaller.

As shown in Figure 6.3, the data with 3-x-3 symmetric structure are generated. The lengths of higher parts are 3 in both sides and the length of lower part is $k-6$ in the middle. The sample size and δ, which is the difference between the two part, affect the finding rate. The value of the higher part (p_1) also affects the result, but not obviously. In the common motif finding problem, e.g. former example: δ=0.4, sample size=14 and p_1=0.95, the finding rate is around 0.80. The finding rate can be increased with scarify of α control. For example, if we use MLT or MCT method with $\alpha = 0.95$, the finding rate is higher than 99.9%.

As shown in Figure 6.4, the data with 5-x-3 asymmetric structure are generated. The higher parts are 5 and 3 while the length of lower part is $k-8$ in the middle. The length of the sequences k also affects the result. The differences between symmetric and asymmetric are very small.

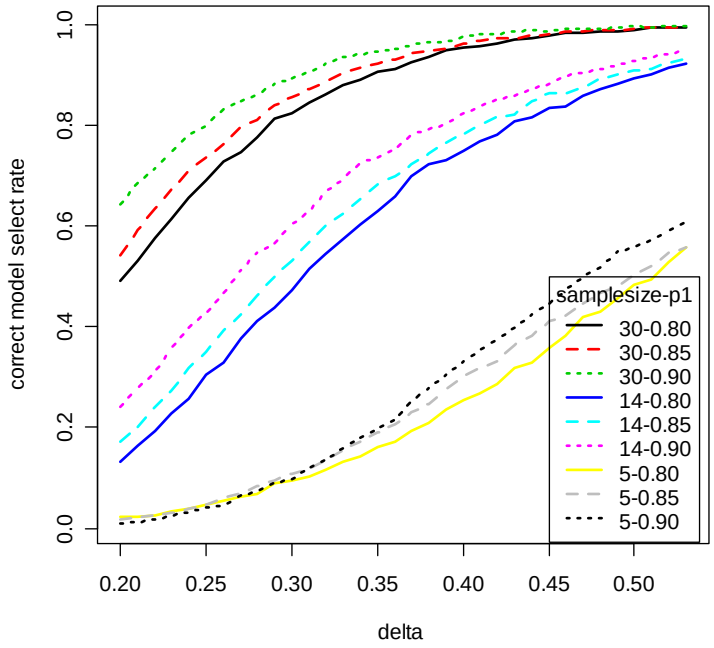

Figure 6.3: Under the alternative,asymmetric 3x3 pattern: power of NIC

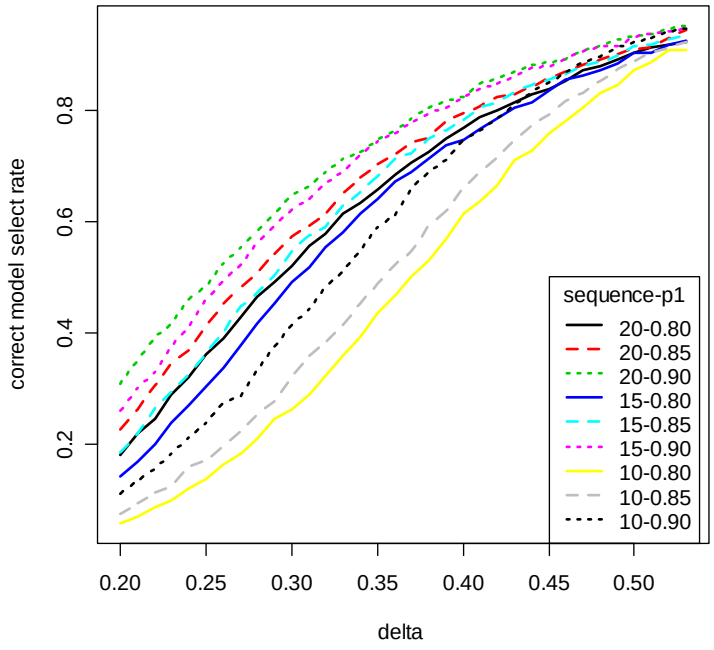

Figure 6.4: Under the alternative, asymmetric 5x3 pattern: power of NIC

Alternatives	Methods	$Asy.H_0$	H_0	H_A^1	H_A^2	H_A^3
0.4/0.4/0.4	MLT	0.9500	0.9600	0.0100	0.0180	0.0120
0.4/0.4/0.4	MCT	0.9500	0.9600	0.0100	0.0190	0.0110
0.4/0.4/0.4	ORIC-lMLE	-	0.4850‡	0.0980	0.0910	0.3260 ‡
0.4/0.4/0.4	MHIC	0.7414	0.7160	0.1190	0.1010	0.0640
0.6/0.6/0.6	MLT	0.9500	0.9600	0.0160	0.0120	0.0120
0.6/0.6/0.6	MCT	0.9500	0.9610	0.0170	0.0110	0.0110
0.6/0.6/0.6	ORIC-lMLE	-	0.5200‡	0.0910	0.0930	0.2960 ‡
0.6/0.6/0.6	MHIC	0.7415	0.7360	0.0900	0.0910	0.0830
0.4/0.6/0.6	MLT	0.0708	0.0890	0.7370	0.0060	0.1680
0.4/0.6/0.6	MCT	0.0708	0.0890	0.7380	0.0060	0.1670
0.4/0.6/0.6	ORIC-lMLE	-	0.0010	0.5680	0.0130	0.4180 ‡
0.4/0.6/0.6	MHIC	0.0073	0.0020	0.8110	0.0090	0.1780
0.4/0.4/0.6	MLT	0.0707	0.0880	0.0080	0.7490	0.1550
0.4/0.4/0.6	MCT	0.0707	0.0880	0.0080	0.7490	0.1550
0.4/0.4/0.6	ORIC-lMLE	-	0.0060	0.0130	0.5860	0.3950 ‡
0.4/0.4/0.6	MHIC	0.0073	0.0070	0.0100	0.8170	0.1660
0.4/0.5/0.6	MLT	0.1319	0.1330	0.1960	0.1900	0.4810
0.4/0.5/0.6	MCT	0.1319	0.1330	0.1950	0.1910	0.4810
0.4/0.5/0.6	ORIC-lMLE	-	0.0200	0.2210	0.2330	0.5260
0.4/0.5/0.6	MHIC	0.0170	0.0200	0.2360	0.2240	0.5200

Table 6.5: 1000 random binomial data for $k = 3$, proportions $p_0 = ... = p_{j-1} = 0.4$,$p_j = 0.4, 0.5, 0.6$, $p_{j+1} = ... = p_k = 0.6$, and sample size n_i is 100.

6.2.3 Simple-order

In this section, we will give further examples of the correct model selection rate over different alternatives under Simple-order restriction. In these simulations, we generate 10000 random binomial data for $k = 3$ and 4 with isotonic means $0.4 \leq p_0 \leq ... \leq p_k \leq 0.4 + \Delta = 0.6$. The sample sizes for $k = 3$ are 100 (Table 6.5). The sample sizes for $k = 4$ are 100 (Table 6.6) and 10 (Table 6.7). When $k = 4$ the behavior of alternative H_A^6 is strange for both sample sizes. We could also take H_A^6 out to achieve a higher classification rate.

6.2.4 Simple-tree order

For asymptotic normal case, we generate 10000 random binomial data for $k = 5$ with means $p_0 = ... = p_{j-1} = .4$, $p_j = ... = p_k = .4 + \Delta = .6$, and the sample size is 5. This simulated data is totally the same as what we generate in previous

Alternatives	Methods	$Asy.H_0$	H_0	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5	H_A^6	H_A^7
H_0										
.4/.4/.4/.4	MLT	.951	.950	.011	.012	.011	.004	.004	.003	.005
.4/.4/.4/.4	MCT	.951	.949	.010	.011	.011	.004	.005	.005	.005
.4/.4/.4/.4	ORIC-lMLE	.672	.672	.102	.091	.110	.000	.005	.014	.005
.4/.4/.4/.4	MHIC	.672	.669	.086	.079	.094	.012	.021	.016	.023
H_0										
.6/.6/.6/.6	MLT	.950	.952	.011	.010	.012	.002	.005	.004	.004
.6/.6/.6/.6	MCT	.950	.952	.010	.009	.011	.002	.005	.006	.004
.6/.6/.6/.6	ORIC-lMLE	.672	.680	.104	.088	.105	.000	.006	.013	.004
.6/.6/.6/.6	MHIC	.672	.676	.089	.077	.091	.011	.023	.017	.017
H_A^1										
.4/.6/.6/.6	MLT	.060	.067	.655	.012	.002	.022	.002	.051	.189
.4/.6/.6/.6	MCT	.060	.063	.629	.009	.001	.018	.002	.082	.197
.4/.6/.6/.6	ORIC-lMLE	.003	.004	.663	.012	.002	.015	.002	.151	.151
.4/.6/.6/.6	MHIC	.003	.003	.702	.015	.003	.023	.002	.054	.197
H_A^2										
.4/.4/.6/.6	MLT	.018	.020	.001	.771	.001	.100	.052	.004	.051
.4/.4/.6/.6	MCT	.018	.019	.001	.736	.001	.102	.069	.004	.067
.4/.4/.6/.6	ORIC-lMLE	.001	.001	.003	.612	.004	.061	.154	.007	.158
.4/.4/.6/.6	MHIC	.001	.001	.002	.785	.002	.101	.053	.004	.053
H_A^3										
.4/.4/.4/.6	MLT	.060	.068	.002	.011	.653	.021	.193	.053	.000
.4/.4/.4/.6	MCT	.060	.065	.000	.009	.627	.018	.198	.082	.001
.4/.4/.4/.6	ORIC-lMLE	.003	.005	.002	.013	.664	.018	.145	.152	.001
.4/.4/.4/.6	MHIC	.003	.005	.004	.014	.696	.022	.201	.056	.001
H_A^4										
.4/.46/.53/.6	MLT	.111	.121	.046	.174	.051	.226	.123	.138	.120
.4/.46/.53/.6	MCT	.111	.114	.036	.154	.044	.212	.134	.174	.132
.4/.46/.53/.6	ORIC-lMLE	.008	.009	.072	.181	.074	**.136**	.147	.233	.148
.4/.46/.53/.6	MHIC	.008	.008	.061	.204	.066	.237	.137	.152	.135
H_A^5										
.4/.4/.46/.6	MLT	.074	.083	.002	.110	.241	.110	.377	.066	.009
.4/.4/.46/.6	MCT	.074	.081	.002	.091	.225	.101	.401	.089	.010
.4/.4/.46/.6	ORIC-lMLE	.004	.005	.006	.098	.293	.072	**.368**	.141	.016
.4/.4/.46/.6	MHIC	.004	.005	.004	.127	.265	.118	.396	.072	.011
H_A^6										
.4/.5/.5/.6	MLT	.152	.167	.131	.036	.129	.128	.076	.264	.070
.4/.5/.5/.6	MCT	.152	.152	.096	.031	.095	.114	.078	**.361**	.074
.4/.5/.5/.6	ORIC-lMLE	.013	.014	.130	.056	.134	.076	.064	**.464**	.062
.4/.5/.5/.6	MHIC	.013	.014	.168	.048	.169	.141	.086	.294	.080
H_A^7										
.4/.53/.6/.6	MLT	.074	.085	.245	.115	.003	.109	.010	.069	.365
.4/.53/.6/.6	MCT	.074	.082	.232	.096	.002	.101	.011	.089	.386
.4/.53/.6/.6	ORIC-lMLE	.004	.005	.298	.100	.006	.072	.016	.142	.361
.4/.53/.6/.6	MHIC	.004	.004	.273	.132	.005	.115	.012	.074	.384

Table 6.6: 1000 random binomial data for $k = 4$, and sample size n_i is 100.

Alternatives	Methods	$Asy.H_0$	H_0	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5	H_A^6	H_A^7
H_0										
.4/.4/.4/.4	MLT	.950	.949	.017	.009	.007	.002	.005	.007	.004
.4/.4/.4/.4	MCT	.950	.950	.016	.008	.005	.001	.006	.010	.004
.4/.4/.4/.4	MHIC	.672	.685	.084	.068	.088	.009	.021	.019	.026
H_0										
.6/.6/.6/.6	MLT	.950	.944	.011	.010	.014	.006	.009	.003	.003
.6/.6/.6/.6	MCT	.950	.948	.006	.010	.011	.004	.010	.007	.004
.6/.6/.6/.6	MHIC	.672	.668	.091	.073	.084	.020	.026	.018	.020
H_1										
.4/.6/.6/.6	MLT	.768	.798	.094	.025	.019	.014	.007	.015	.028
.4/.6/.6/.6	MCT	.768	.794	.084	.019	.013	.012	.009	.039	.030
.4/.6/.6/.6	MHIC	.324	.329	.314	.096	.080	.034	.020	.045	.082
H_2										
.4/.4/.6/.6	MLT	.706	.682	.025	.162	.027	.026	.034	.009	.035
.4/.4/.6/.6	MCT	.706	.692	.017	.147	.021	.027	.039	.018	.039
.4/.4/.6/.6	MHIC	.259	.250	.092	.346	.093	.044	.072	.034	.069
H_3										
.4/.4/.4/.6	MLT	.768	.761	.018	.024	.116	.019	.036	.017	.009
.4/.4/.4/.6	MCT	.768	.761	.012	.021	.096	.017	.042	.042	.009
.4/.4/.4/.6	MHIC	.324	.314	.062	.112	.311	.038	.085	.052	.026
H_4										
.4/.46/.53/.6	MLT	.774	.753	.032	.062	.045	.030	.042	.011	.025
.4/.46/.53/.6	MCT	.774	.758	.026	.053	.034	.028	.041	.029	.031
.4/.46/.53/.6	MHIC	.326	.328	.119	.162	.153	.056	.074	.040	.068
H_5										
.4/.4/.46/.6	MLT	.762	.780	.013	.066	.071	.013	.033	.013	.011
.4/.4/.46/.6	MCT	.762	.787	.008	.056	.061	.015	.035	.027	.011
.4/.4/.46/.6	MHIC	.315	.311	.070	.182	.232	.031	.090	.059	.025
H_6										
.4/.5/.5/.6	MLT	.794	.784	.046	.036	.058	.023	.013	.022	.018
.4/.5/.5/.6	MCT	.794	.783	.034	.034	.041	.020	.017	.055	.016
.4/.5/.5/.6	MHIC	.350	.381	.159	.113	.173	.036	.039	.057	.042
H_7										
.4/.53/.6/.6	MLT	.762	.760	.075	.057	.020	.021	.013	.017	.037
.4/.53/.6/.6	MCT	.762	.762	.066	.051	.009	.020	.014	.038	.040
.4/.53/.6/.6	MHIC	.315	.320	.236	.164	.068	.037	.034	.048	.093

Table 6.7: 1000 random binomial data for $k = 4$, and sample size n_i is 10.

Alternatives	Meth.	j	$Asy.H_0$	H_0	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5
.4/.4/.4/.4/.4/.4	MLT	-	.9501	.9440	.0080	.0200	.0105	.0105	.0070
.4/.4/.4/.4/.4/.4	MCT	-	.9501	.9505	.0084	.0191	.0090	.0080	.0050
.4/.4/.4/.4/.4/.4	ORIC-lMLE	-	.5861	.5470	.0800	.1053	.0754	.0882	.1041
.6/.6/.6/.6/.6/.6	MLT	-	.9501	.9429	.0120	.0101	.0118	.0119	.0113
.6/.6/.6/.6/.6/.6	MCT	-	.9501	.9476	.0106	.0094	.0109	.0109	.0106
.6/.6/.6/.6/.6/.6	ORIC-lMLE	-	.5862	.5568	.0784	.0807	.0906	.0890	.1045
.4/.6/.6/.6/.6/.6	MLT	1-5	.2303	.2425	.1299	.1327	.1554	.1611	.1784
.4/.6/.6/.6/.6/.6	MCT	1-5	.2303	.2499	.1280	.1321	.1541	.1593	.1766
.4/.6/.6/.6/.6/.6	ORIC-lMLE	1-5	.0131	.0094	.1666	.1730	.1972	.2121	.2417
.4/.4/.6/.6/.6/.6	MLT	2-5	.2690	.2930	.0003	.1647	.1660	.1812	.1948
.4/.4/.6/.6/.6/.6	MCT	2-5	.2690	.3136	.0003	.1596	.1608	.1752	.1905
.4/.4/.6/.6/.6/.6	ORIC-lMLE	2-5	.0190	.0157	.0005	.2203	.2299	.2520	.2816
.4/.4/.4/.6/.6/.6	MLT	3-5	.3247	.3484	.0001	.0003	.1959	.2091	.2462
.4/.4/.4/.6/.6/.6	MCT	3-5	.3247	.3703	.0001	.0003	.1877	.2030	.2386
.4/.4/.4/.6/.6/.6	ORIC-lMLE	3-5	.0302	.0222	.0004	.0006	.2860	.3098	.3810
.4/.4/.4/.4/.6/.6	MLT	4-5	.4119	.4487	.0004	.0007	.0009	.2630	.2863
.4/.4/.4/.4/.6/.6	MCT	4-5	.4119	.4632	.0004	.0007	.0009	.2557	.2791
.4/.4/.4/.4/.6/.6	ORIC-lMLE	4-5	.0560	.0502	.0016	.0023	.0028	.4377	.5054
.4/.4/.4/.4/.4/.6	MLT	5	.5712	.6010	.0026	.0016	.0020	.0020	.3908
.4/.4/.4/.4/.4/.6	MCT	5	.5712	.5921	.0023	.0017	.0021	.0018	.4000
.4/.4/.4/.4/.4/.6	ORIC-lMLE	5	.1373	.1232	.0112	.0100	.0106	.0097	.8353

Table 6.8: 10000 random binomial data for $k = 5$, proportions $p_0 = ... = p_{j-1} = .4$, $p_j = ... = p_k = .6$, sample size n_i is 50.

section(see Table 6.1). But the models are different. We want to find out if there is any treatment which is different from the control. The result is shown in Table 6.9. As shown in the table, the correct model selection rates vary from different patterns. This result also verifies the theoretical value of the asymptotic power and correct model selection rate. MHIC is not considered here, because it is totally identical with ORIC-lMLE under Simple-tree order restriction.

6.3 Conclusion

The simulation studies we have done in last section, verify that "no uniformly powerful test exits" under order restriction and "no uniformly powerful model selection method exits". MCT and MLT have a higher correct model selection rate (CR) to detect the Change-point than ORIC-lMLE. ORIC-lMLE is very good to detect Simple-order, however the misclassification rate (MR) for ORIC-lMLE is also high

Alternatives	Meth.	$Asy.H_0$	H_0	H_A^1	H_A^2	H_A^3
0.3/0.3/0.3	MLT	-	0.9438	0.0226	0.0204	0.0132
0.3/0.3/0.3	MCT	-	0.9456	0.0233	0.0176	0.0135
0.3/0.3/0.3	ORIC-lMLE	-	0.5012	0.0882	0.0922	0.3184
0.3/0.3/0.3	MHIC	-	0.7264	0.1050	0.1145	0.0541
0.7/0.7/0.7	MLT	-	0.9514	0.0158	0.0194	0.0134
0.7/0.7/0.7	MCT	-	0.9527	0.0140	0.0201	0.0132
0.7/0.7/0.7	ORIC-lMLE	-	0.5155	0.0849	0.0899	0.3097
0.7/0.7/0.7	MHIC	-	0.7406	0.1023	0.1049	0.0522
0.3/0.6/0.7	MLT	-	0.0039	0.5308	0.0120	0.4533
0.3/0.6/0.7	MCT	-	0.0039	0.5327	0.0118	0.4516
0.3/0.6/0.7	ORIC-lMLE	-	0.0002	0.3472	0.0115	0.6411
0.3/0.6/0.7	MHIC	-	0.0003	0.5334	0.0123	0.4540
0.3/0.5/0.7	MLT	-	0.0087	0.1399	0.1471	0.7043
0.3/0.5/0.7	MCT	-	0.0087	0.1401	0.1470	0.7042
0.3/0.5/0.7	ORIC-lMLE	-	0.0005	0.0993	0.1024	0.7978
0.3/0.5/0.7	MHIC	-	0.0005	0.1421	0.1498	0.7076
0.3/0.4/0.7	MLT	-	0.0038	0.0124	0.5323	0.4515
0.3/0.4/0.7	MCT	-	0.0038	0.0122	0.5354	0.4486
0.3/0.4/0.7	ORIC-lMLE	-	0.0002	0.0122	0.3438	0.6438
0.3/0.4/0.7	MHIC	-	0.0002	0.0125	0.5348	0.4525
0.3/0.4/0.6	MLT	-	0.0849	0.0696	0.3759	0.4696
0.3/0.4/0.6	MCT	-	0.0853	0.0688	0.3839	0.4620
0.3/0.4/0.6	ORIC-lMLE	-	0.0099	0.1060	0.3447	0.5394
0.3/0.4/0.6	MHIC	-	0.0088	0.0819	0.4138	0.4955
0.3/0.4/0.5	MLT	-	0.3705	0.1663	0.1617	0.3015
0.3/0.4/0.5	MCT	-	0.3721	0.1637	0.1656	0.2986
0.3/0.4/0.5	ORIC-lMLE	-	0.0886	0.3290	0.2595	0.3229
0.3/0.4/0.5	MHIC	-	0.0864	0.2654	0.2497	0.3985
0.3/0.35/0.35	MLT	-	0.8678	0.0651	0.0288	0.0383
0.3/0.35/0.35	MCT	-	0.8679	0.0649	0.0289	0.0383
0.3/0.35/0.35	ORIC-lMLE	-	0.4311	0.2040	0.1165	0.2484
0.3/0.35/0.35	MHIC	-	0.5505	0.2241	0.1165	0.1089
0.3/0.4/0.4	MLT	-	0.7291	0.1614	0.0352	0.0743
0.3/0.4/0.4	MCT	-	0.7292	0.1614	0.0354	0.0740
0.3/0.4/0.4	ORIC-lMLE	-	0.2946	0.3483	0.1261	0.2310
0.3/0.4/0.4	MHIC	-	0.3480	0.3897	0.0999	0.1624
0.3/0.5/0.5	MLT	-	0.2947	0.5158	0.0183	0.1712
0.3/0.5/0.5	MCT	-	0.2947	0.5174	0.0187	0.1692
0.3/0.5/0.5	ORIC-lMLE	-	0.0551	0.5079	0.0673	0.3697
0.3/0.5/0.5	MHIC	-	0.0634	0.6810	0.0360	0.2196
0.3/0.6/0.6	MLT	-	0.0009	0.9269	0.0000	0.0722
0.3/0.6/0.6	MCT	-	0.0009	0.9326	0.0000	0.0665
0.3/0.6/0.6	ORIC-lMLE	-	0.0002	0.6001	0.0000	0.3997
0.3/0.6/0.6	MHIC	-	0.0002	0.9275	0.0000	0.0723

Table 6.9: 10000 random binomial data for $k = 3$, with different none center parameters. Sample size is 50.

when the true model is not the Simple-order.

From the simulation, we also find that MLT has a very similar power behavior as MCT, under asymptotic normality. Under Simple-order, Simple-tree and Single Change-point order restriction, there is a one-to-one correspondence between MLT and MCT, i.e. under these tree order restrictions, for each MCT method we could find a corresponded MLT whose power is no less than the MCT.

Here we recommend our readers to use MLT which has an "empirical average power" too. Furthermore, from our simulation we see that MLT is no worse than MCT and MLT is a test-based model selection method.

Chapter 7

Software

7.1 Multivariate Normal Distribution and package Mvnorm

As seen in last chapter, the distribution functions of multivariate normal (mvn) which is the fundamental evaluations of the test statistics studied over the whole thesis, need to be calculated. In the first section of this chapter, three algorithms for calculating mvn are introduced and compared. A short and brief code for how to use our package Biotrend, will also be given.

Miwa et al. (2003) proposed an numerical algorithm for evaluating multivariate normal probabilities. Starting with version 0.9-0 of the **mvtnorm** package (Hothorn et al., 2001), this algorithm is available to the R community. In this section we will give a brief introduction to Miwa's procedure and compare it to a quasi-randomized Monte-Carlo procedure proposed by Genz and Bretz (1999), which has been available through **mvtnorm** for some years now, both with respect to computing time and accuracy. Craig (2008) made an improvement in Miwa's algorithm and proposed the Auto Regression (AR) and Moving Average (MA) model to describe the structure of the correlation matrix. The Miwa's algorithm and Craig's algorithm are both applicable to problems with dimension smaller than 20, whereas the pro-

cedures by Genz and Bretz (1999) can be used to evaluate 1000-dimensional normal distributions. At the end of this section, a suggestion is given for choosing a suitable algorithm in different situations.

All the codes and software that mentioned in this chapter are included in the CD-rom.

7.1.1 Definition and properties

The important role of mvn distribution played in this thesis, is to calculate the rectangular quartile of the test statistics for LRT and MCT. In the next section, we will also transform the q-variate chi-square distribution to mvn distribution. Many literatures have discussed about mvn distribution and the definition from Srivastava and Carter (1975), Bretz (1999) and Miwa et al. (2003) will be used.

Let random variables $Y = \{y_1, ..., y_k\}$ be i.i.d standard normal distributed, i.e. $N(U_Y, \sigma_Y^2$. A random variable vector $X = \{x_1, ..., x_k\}$ is called k-variate mvn distribution $N_k(U, \Sigma)$ if X has the same distribution as $U + BY$, where B is a k by k matrix such as $BB^t = \Sigma$ and $U = \{u_1, ..., u_k\}$ is a vector of constants. U is called the mean of X. We denote the covariance matrix $\Sigma = \{\sigma_{ij}\}, 1 \leq i, j \leq k$ and $R = \{\rho_{ij}\} = \{\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\}$ is the correlation matrix of X.

The covariance matrix Σ has many special structures. The distribution of X is called singular if the determinant of Σ, noted as $|\Sigma|$, is equal to zero and is called non-singular if the value of $|\Sigma|$ is bigger than zero. Since singular mvn distribution can be transformed into non-singular mvn distribution, here only the non-singular situation is considered. In special case with $k = 1$, $U = 0$ and $\Sigma = \{1\}$, $N_1(0, 1)$ is the traditional univariate standard norm distribution.

For given mean U and covariance matrix Σ, the probability density function of X is

$$\phi_k(X; U, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|} \exp\{-\frac{1}{2}(X - U)^t \Sigma^{-1}(X - U)\} \tag{7.1}$$

The problem for calculating any non-centered orthant probability of a non-singular

multivariate normal distribution is described by Miwa et al. (2003).

$$\begin{aligned} P_k(U,\Sigma) &= Pr\{x_i \geq 0;\, 1 \leq i \leq k\} \\ &= \int_0^\infty ... \int_0^\infty \phi_k(X;U,\Sigma)\, dx_1 ... dx_k \end{aligned} \tag{7.2}$$

For given upper limit $D = \{d_1, ..., d_k\}$, the one-sided normal distribution function can be expressed as a non-centered orthant probability

$$\begin{aligned} \Phi_k(D) &= Pr\{x_i \leq d_i;\, 1 \leq i \leq k\} \\ &= Pr\{-x_i \geq -d_i;\, 1 \leq i \leq k\} \\ &= P_k(-U + D, \Sigma) \\ &= \int_0^\infty ... \int_0^\infty \phi_k(X;-U+D,\Sigma)\, dx_1 ... dx_k \\ &= \int_{-\infty}^{d_1} ... \int_{-\infty}^{d_k} \phi_k(X;U,\Sigma)\, dx_1 ... dx_k \end{aligned} \tag{7.3}$$

One special case of the orthant probability is the orthoscheme probability. An orthant probability $P_k(U,R)$ is called orthoscheme probability, if the correlation matrix R is a tridiagonal matrix. More details of orthoscheme probability will be given in Section 7.1.3.

The two-sided probability, with $D = \{d_1, ..., d_k\}$ and $E = \{e_1, ..., e_k\}$ as the upper and lower limits, is defined as following

$$\begin{aligned} \Phi_k(E,D) &= Pr\{e_i \leq x_i \leq d_i;\, 1 \leq i \leq k\} \\ &= \int_{e_1}^{d_1} ... \int_{e_k}^{d_k} \phi_k(X;U,\Sigma)\, dx_1 ... dx_k \end{aligned} \tag{7.4}$$

and can be calculated from 2^k k-dimensional one-sided probabilities which have the

same mean and covariance matrix.

$$\begin{aligned}\Phi_k(E,D) &= Pr\{e_i \le x_i \le d_i, 1 \le i \le k\} \\ &= Pr\{x_i \le d_i, 1 \le i \le k\} \\ &- \sum_{j=1}^{k}(Pr\{x_1 \le d_1, ..., x_j \le e_j, ..., x_k \le d_k\} \\ &+ \sum_{j=1}^{k}\sum_{h=j}^{k}(Pr\{x_1 \le d_1, ..., x_j \le e_j, ..., x_h \le e_h, ..., x_k \le d_k\} \\ &- \sum_{j=1}^{k}\sum_{h=j}^{k}\sum_{l=h}^{k}... \end{aligned} \tag{7.5}$$

Miwa et al. (2003) provided a numerical algorithm which is not linear in dimension k. Genz and Bretz (1999) developed a Monte-Carlo procedure which can calculate the two-sided probability in linear time. In the following chapters, for simplicity, we assume that the random vector X has unit variances, the mean $U = \{u_1, ..., u_k\} = \{0, ..., 0\}$ and the covariance matrix Σ is equal to the correlation matrix R.

7.1.2 Monte-Carlo algorithm

Genz (1992) transformed the k dimensional integral of two-sided probability into a $k-1$ dimensional integral over a hypercube. By doing this, this algorithm avoids the infinite integral bound which is hard to deal with. Furthermore, many efficient numerical integral algorithms can be applied in this hypercube region. Genz and Bretz (1999) gave further improvement of this algorithm.

The multi-dimensional two-sided probability of X with zero mean is

$$\begin{aligned}\Phi_k(E,D) &= Pr\{e_i \le x_i \le d_i;\ 1 \le i \le k\} \\ &= \int_{e_1}^{d_1}\int_{e_2}^{d_2}...\int_{e_k}^{d_k}\frac{1}{(2\pi)^{k/2}|\Sigma|}\exp\{-\frac{1}{2}X^t\Sigma^{-1}X\}dX \end{aligned} \tag{7.6}$$

here, $D = \{d_1, ..., d_k\}$ and $E = \{e_1, ..., e_k\}$ are the upper and lower limits of the integral region. The correlation matrix Σ, which is mentioned in last section, can

be decomposed into BB^t. B is the Cholesky triangle, which is a lower triangular matrix. Note $Y = \{y_1, ..., y_k\}$ as the decomposition vector and $X = BY$. This implies $X^t\Sigma^{-1}X = Y^tB^tB^{-t}B^{-1}BY = Y^tY$ and $dX = |B|dY = |\Sigma^{-1/2}|dY$. By knowing the integral region of X as $E \leq X = BY \leq D$, we can calculate the integral region (E', D') of Y, where $E' = \{e'_1, ..., e'_k\}$ and $D' = \{d'_1, ..., d'_k\}$ are the upper and lower limits of the integral region for Y

$$(E' \leq Y \leq D') = \begin{cases} \dfrac{e_i}{b_{ii}} \leq y_i \leq \dfrac{d_i}{b_{ii}}, & i = 1 \\ \dfrac{e_i - \sum_{j=1}^{i-1} b_{ij}y_j}{b_{ii}} \leq y_i \leq \dfrac{d_i - \sum_{j=1}^{i-1} b_{ij}y_j}{b_{ii}}, & \text{i=2,...,k} \end{cases} \tag{7.7}$$

The probability function is transformed to

$$\begin{aligned} \Phi_k(E, D) &= \frac{1}{(2\pi)^{k/2}} \int_{e'_1}^{d'_1} \int_{e'_2(y_1)}^{d'_2(y_1)} ... \int_{e'_k(y_1,...y_{k-1})}^{d'_k(y_1,...y_{k-1})} \exp\{-\frac{1}{2}Y^tY\}dY \\ &= \frac{1}{(2\pi)^{k/2}} \int_{e'_1}^{d'_1} \exp^{-\frac{y_1^2}{2}} \int_{e'_2(y_1)}^{d'_2(y_1)} \exp^{-\frac{y_2^2}{2}} ... \int_{e'_k(y_1,...y_{k-1})}^{d'_k(y_1,...y_{k-1})} \exp^{-\frac{y_k^2}{2}} dY \end{aligned} \tag{7.8}$$

Now random variable X has been successfully separated by Cholesky decomposition. However the integral function of Y is still a complicated exponential function. Transformation is made again to reduce the complicity. The components of Y are independent. They can be transformed separately as $y_i = \Phi_1^{-1}(z_i)$, where

$$\Phi_1(y_i) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{y_i} \exp^{-\frac{v^2}{2}} dv \tag{7.9}$$

has been defined as univariate standard norm distribution in former part of this section and ϕ_1 is the density function of it. The following equation is achieved by differentiating both sides of it

$$dz_i = d\Phi_1(y_i) = d\int_{-\infty}^{y_i} \exp^{-\frac{v^2}{2}} dv = \frac{1}{(2\pi)^{1/2}} \exp^{-\frac{y_i^2}{2}} dy_i \tag{7.10}$$

furthermore,

$$\begin{aligned} dZ &= \{dz_1, ..., dz_k\} \\ &= \{\frac{1}{(2\pi)^{1/2}} \exp^{-\frac{y_1^2}{2}} dy_1, ..., \frac{1}{(2\pi)^{1/2}} \exp^{-\frac{y_k^2}{2}} dy_k\} \end{aligned} \tag{7.11}$$

By putting this into Equation 7.8, they have got

$$\Phi_k(E, D) = \int_{e_1''}^{d_1''} \int_{e_2''(z_1)}^{d_2''(z_1)} ... \int_{e_k''(z_1,...z_{k-1})}^{d_k''(z_1,...z_{k-1})} dZ \tag{7.12}$$

with the integral region of Z

$$(E'' \leq Z \leq D'') = \begin{cases} \Phi_1(e_i/b_{ii}) \leq z_i \leq \Phi_1(d_i/b_{ii}), & i = 1 \\ \Phi_1(\frac{e_i - \sum_{j=1}^{i-1} b_{ij}\Phi_1^{-1}(z_j)}{b_{ii}}) \leq z_i \leq \Phi_1(\frac{d_i - \sum_{j=1}^{i-1} b_{ij}\Phi_1^{-1}(z_j)}{b_{ii}}), & \text{i=2,...,k} \end{cases} \tag{7.13}$$

The integral region of Equation 7.12, is hard to calculate by common numerical integral algorithms. For example, the upper and lower limit of z_k is a function of $\{z_1, ..., z_{k-1}\}$. Assume G is the number of grid points for one variable, we need in total G^{k-1} grid points to evaluated the function of

$$f(z_1, ...z_{k-1}) = \int_{e_k''(z_1,...z_{k-1})}^{d_k''(z_1,...z_{k-1})} dz_k \tag{7.14}$$

An Monte-Carlo algorithm, which can calculate integral within closed region in linear time, is suggested by Bretz (1999). However, special cases, such as orthoscheme probabilities that the integral region of z_k is only a function of z_{k-1}, can be evaluated in linear time by recursive numerical method (Miwa et al., 2003). Detailed discussion of these situations will be given in next section.

Before applying Monte-Carlo integral algorithm, the integral region should be uniformed for easier programming and better error estimation. Genz (1992) transformed the equation again to uniform the integral region by taking $Z = E'' + W(D'' - E'')$.

Because $E'' \leq Z \leq D''$, the integral region of W is $\mathbf{0} \leq W \leq \mathbf{1}$, where $\mathbf{0}$ and $\mathbf{1}$ are k-dimension constant vectors. After transform the original integral for three times, we have the final integral as

$$\begin{aligned}\Phi_k(E, D) &= (d_1''' - e_1''') \int_0^1 (d_2''' - e_2''') \int_0^1 ...(d_k''' - e_k''') \int_0^1 dW \\ &= \int_0^1 \int_0^1 ... \int_0^1 f(w) dW \end{aligned} \tag{7.15}$$

here, $f(w) = (d_k''' - e_k''')(d_k''' - e_k''')...(d_k''' - e_k''')$ with value D''' and E''' which can be described by function of W:

$$e_1''' = \Phi_1(e_1/b_{11}), d_1''' = \Phi_1(d_1/b_{11}), \qquad i = 1$$

$$\begin{aligned} e_i''' &= \frac{e_i - \sum_{j=1}^{i-1} b_{ij} \Phi_1^{-1}(e_j''' + w_j(d_j''' - e_j'''))}{b_{ii}}, \quad \text{i=2,...,k} \\ d_i''' &= \frac{e_i - \sum_{j=1}^{i-1} b_{ij} \Phi_1^{-1}(e_j''' + w_j(d_j''' - e_j'''))}{b_{ii}}, \quad \text{i=2,...,k} \end{aligned} \tag{7.16}$$

This integral can be calculated by the quasi-randomized Monte-Carlo procedure developed by Genz and Bretz (1999) in linear time.

7.1.3 Numerical algorithm

In former sections we have seen that the structures of the correlation matrix have a strong influence to the calculation algorithm of the integral. Before introducing Miwa's algorithm and Craig's algorithm, we first give definition and discussion of the structure. Here also for simplicity, the random vector X is assumed to have unit variances.

Miwa's algorithm

A tridiagonal matrix is a matrix which has nonzero entries only in its main diagonal row, the one above this row and the one below this row. For example, matrix R_1 is a n by n tridiagonal correlation matrix for k-variate mvn random vector $X =$

$\{x_1, ..., x_k\}$:

$$\mathbf{R_1} = \begin{pmatrix} 1 & r_1 & & & & \\ r_1 & 1 & r_2 & & & \\ & r_2 & 1 & r_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & r_{n-1} & 1 & r_n \\ & & & & r_n & 1 \end{pmatrix}$$

Note B_1 as the Choleshy decomposition of R_1. Then we have $R_1 = B_1 B_1^t$, where B have only nonzero entries along the diagonal row and the one below this row

$$\mathbf{B_1} = \begin{pmatrix} 1 & & & & & \\ b_{21} & b_{22} & & & & \\ & b_{32} & b_{33} & & & \\ & & \ddots & \ddots & \ddots & \\ & & & b_{(n-1)(n-2)} & b_{(n-1)(n-1)} & \\ & & & & b_{n(n-1)} & b_{nn} \end{pmatrix}$$

By using transformation matrix B_1, we can transform $X = \{x_1, ..., x_k\}$ into Y that $X = B_1 Y$, where vector $Y = \{y_1, ..., y_k\}$ is i.i.d standard normal distributed as we defined in former chapter. By considering $\{x_1, ..., x_k\}$ as an ordered time series process, we have $x_1 = y_1$ and $x_i = b_{i(i-1)} y_{i-1} + b_{ii} y_i$ for $i > 1$, so X is a time series process which has Moving Average (MA) order one(Craig, 2008). In order to distinguish from the MA process in time series, we note the order of the MA process here as Miwa-MA(1). Miwa et al. (2003) developed a numerical algorithm to calculate the orthoscheme probability, $P_k(U, R_1)$ where R_1 is tridiagonal matrix whose order is Miwa-MA(1) .

Now we give a former definition for Miwa-MA model. Let R_b be a k by k correlation matrix and its Cholesky decomposition matrix is B that $R_b = BB^t$. A k-variate mvn random vector X can be transformed into Y, which is a vector of k i.i.d normal variables, by $X = BY$. Sequence $X = \{X_1, ...X_k\}$ with correlation matrix R_b is

called Moving Average (MA) order r or in short Miwa-MA(r) process, when

$$\begin{aligned} x_i &= \sum_{j=1}^{i} b_{ij} y_j, \;\; i \leq r \\ x_i &= \sum_{j=i-r}^{i} b_{ij} y_j, \;\; i > r \end{aligned} \tag{7.17}$$

I.e, random vector X_b with a 4 by 4 correlation matrix

$$\mathbf{R_b} = \begin{pmatrix} 1 & .5 & 0 & 0 \\ .5 & 1 & .5 & 0 \\ 0 & .5 & 1 & .5 \\ 0 & 0 & .5 & 1 \end{pmatrix}$$

is a Miwa-MA(1) process. Here we have $X_b = BY$, where B is the decompose matrix of R_b that $R_a = BB^t$ and Y is k-vector of i.i.d standard norm distribution.

Slightly different from Miwa's procedure, we apply the results from last section to introduce the algorithm which calculate the orthoscheme probablities and show that Miwa's algorithm and Genz/Bretz' algorithm only have difference in the last integral calculation step for calculating orthoscheme probability. Let $E = -U$,$D = \{\infty, ..., \infty\}$ and tridiagonal correlation matrix R_1, the orthoscheme probability $P_k(U, R_1)$ can be calculate from two-sided probability $\Phi_k(E, D)$ with zero mean and same correlation matrix R_1 by

$$\begin{aligned} P_k(U, R_1) &= \Phi_k(E, D) \\ &= \int_{-u_1}^{\infty} ... \int_{-u_k}^{\infty} \phi_k(X; 0, R_1)\, dx_1 ... dx_k \end{aligned} \tag{7.18}$$

by applying Equation 7.8, we have

$$\begin{aligned}
&P_k(U, R_1) \\
&= \int_{-u_1}^{\infty} ... \int_{-u_k}^{\infty} \phi_k(X; 0, R_1)\, dx_1 ... dx_k \\
&= \frac{1}{(2\pi)^{k/2}} \int_{e_1'}^{\infty} \exp^{-\frac{y_1^2}{2}} \int_{e_2'(y_1)}^{\infty} \exp^{-\frac{y_2^2}{2}} ... \int_{e_k'(y_{k-1})}^{\infty} \exp^{-\frac{y_k^2}{2}}\, dY \\
&= \frac{1}{(2\pi)^{k/2}} \int_{-u_1}^{\infty} \exp^{-\frac{y_1^2}{2}} \int_{(-u_2 - b_{21} y_1)/b_{22}}^{\infty} \exp^{-\frac{y_2^2}{2}} ... \int_{(-u_k - b_{k(k-1)} y_{k-1})/b_{kk}}^{\infty} \exp^{-\frac{y_k^2}{2}}\, dY
\end{aligned} \tag{7.19}$$

Miwa et al. (2003) introduced a recursive computational approach which can calculate this integral in linear time. He defined

$$\begin{aligned}
f_{k-1}(y) &= \frac{1}{(2\pi)^{1/2}} \int_{(-u_k - b_{k(k-1)} y)/b_{kk}}^{\infty} \exp^{-\frac{v^2}{2}}\, dv \\
f_{i-1}(y) &= \frac{1}{(2\pi)^{1/2}} \int_{(-u_i - b_{i(i-1)} y)/b_{ii}}^{\infty} f_i(v) \exp^{-\frac{v^2}{2}}\, dv, \quad 2 \le i \le k-1,
\end{aligned} \tag{7.20}$$

so that the required probability is transformed into

$$P_k(U, R_1) = \frac{1}{(2\pi)^{1/2}} \int_{-u_1}^{\infty} f_1(v) \exp^{-\frac{v^2}{2}}\, dv \tag{7.21}$$

Then $f_{k-1}(y)$ is calculated first over an optimal designed grid points. The value of f_{k-1} is stored in an array which is correspond to the grid points. By using Equation 7.29, $f_{k-2}, f_{k-3}, ...$ are calculated consequently. So this algorithm is linear algorithm for arbitrary value of k and sufficient accuracy can be achieved by increasing the number of the grid point.

Here we also improve this algorithm by using the results from former section. Apply Equation 7.8 and Equation 7.12, the probability is

$$P_k(U, R_1) = \int_{-u_1}^{\infty} \int_{(-u_2 - b_{21} \Phi_1^{-1}(z_1))/b_{22}}^{\infty} ... \int_{(-u_k - b_{k(k-1)} \Phi_1^{-1}(z_{k-1}))/b_{kk}}^{\infty} dZ \tag{7.22}$$

Similar as Miwa, we define

$$f'_{k-1}(z) = \int_{(-u_k - b_{k(k-1)}\Phi_1^{-1}(z))/b_{kk}}^{\infty} dv$$
$$f'_{i-1}(z) = \int_{(-u_i - b_{i(i-1)}\Phi_1^{-1}(z))/b_{ii}}^{\infty} f'_i(v)dv, 2 \leq i \leq k-1, \tag{7.23}$$

so that the required probability is transformed into

$$P_k(U, R_1) = \int_{-u_1}^{\infty} f'_1(v)dv \tag{7.24}$$

By saving the time from multiplication and making the value of the standard normal distribution function into pre-calculated tables, this algorithm is faster than the former one.

Simply to show, algorithm for two-side probabilities with tridiagonal correlation matrix is also available, such that we can transforms the probabilities into

$$\Phi(E, D) = \int_{e_1}^{e_1} \int_{(e_1 - b_{21}\Phi_1^{-1}(z_1))/b_{22}}^{(d_1 - b_{21}\Phi_1^{-1}(z_1))/b_{22}} \ldots \int_{(e_k - b_{k(k-1)}\Phi_1^{-1}(z_{k-1}))/b_{kk}}^{(d_k - b_{k(k-1)}\Phi_1^{-1}(z_{k-1}))/b_{kk}} dZ \tag{7.25}$$

For probabilities, which are not orthoscheme, Miwa et al. (2003) also gave an algorithm which transform the probabilities into sum of many orthoscheme probabilities. The total number of the orthoscheme probabilities depends on the complexity of the correlation matrix, which can be measured by orthoscheme order. Miwa et al. (2003) has given a definition of it:

"For $0 \leq r \leq m-3$ an m $\times$ m symmetric matrix $\Sigma = \{\sigma_{ij}\}$ is said to have 'orthoscheme order r' if $\sigma_{ij} = 0$ for $1 \leq i \leq r, i+2 \leq j \leq m$ and $\sigma_{r+1,j} \neq 0$ for some j satisfying $r+3 \leq j \leq m$. A tridiagonal matrix is defined to have orthoscheme order $m-2$."

For example, for 6-dimension correlation matrix

$$\mathbf{R_{6,6}} = \begin{pmatrix} 1 & \times & 0 & 0 & 0 & 0 \\ \times & 1 & \times & \times & \times & \times \\ 0 & \times & 1 & \times & \times & \times \\ 0 & \times & \times & 1 & \times & \times \\ 0 & \times & \times & \times & 1 & \times \\ 0 & \times & \times & \times & \times & 1 \end{pmatrix}$$

here $\times$ represents the nonzero entries, the orthoscheme order is one. However, this definition has a disadvantage that the orthoscheme order may be changed by per permutation. I.e. if we interchange the $2nd$ and $4th$ rows of $R_{6,6}$, a new correlation matrix

$$\mathbf{R'_{6,6}} = \begin{pmatrix} 1 & 0 & 0 & 0 & \times & 0 \\ 0 & 1 & \times & \times & \times & \times \\ 0 & \times & 1 & \times & \times & \times \\ 0 & \times & \times & 1 & \times & \times \\ \times & \times & \times & \times & 1 & \times \\ 0 & \times & \times & \times & \times & 1 \end{pmatrix}$$

has orthoscheme order zero. We left the problem, that how to permutate a matrix properly, as an open question and assumed that the matrix have already been properly permutated.

Let R_r be any k by k correlation matrix which has orthoscheme order r, here $0 \leq r \leq k-3$. Then for any given mean vector U, the one-sided orthant probability can be calculated from at most $k-r-1$ one-sided orthant probabilities whose correlation matrices have orthemscheme order larger than r. By repeating this procedure, the one-sided orthant probability can be calculated from at most $(k-1)!$ orthemscheme probabilities. The algorithm for calculating such probabilities is a recursive linear integration procedure. The total order of a one-sided problem is $G \times k!$, where G is the number of grid points for integration. The total order of this two-sided problem is $G \times k! \times 2^k$, where G is the number of grid points for integration. For more detail

of this procedure, please refer to the paper of Miwa et al. (2003).

Craig's algorithm

By using the duality between Moving Average model and Auto Regression (AR) model, we can introduce AR model which is another important model in time series. For correlation matrix R_a that the reverse of R_a is tridiagonal, we can decompose R_a into $R_a = A^{-t}A^{-1}$, where A is a Cholesky decomposition matrix. Then we have $AX = Y$. By considering $\{x_1, ..., x_k\}$ as an ordered time series process, we have $x_1 = y_1$ and $a_{ii}x_i = -a_{i(i-1)}x_{i-1} + y_i$ for $i > 1$, so X is a time series process which has AR order one. In order to distinguish from the AR process in time series, we note the order of the AR process here as Craig-AR(1). A numerical algorithm is also developed to calculate this edge orthoscheme probability $P_k(U, R_a)$, where R_a is reverse tridiagonal matrix whose order is Craig-AR(1) (Craig, 2008).

Now we give a former definition for Craig-AR model. Let R_a be a k by k correlation matrix and the Cholesky decomposition matrix for its reverse matrix is A that $R_a^{-1} = AA^t$. A k-variate mvn random vector X can be transformed into Y, which is a vector of k i.i.d normal variables, by $AX = Y$. Sequence $X = \{X_1, ...X_k\}$ with correlation matrix R_a^{-1} is called AR order r or in short Craig-AR(r) process, when

$$\begin{aligned} a_{ii}x_i &= \sum_{j=1}^{i-1} a_{ij}x_j + y_i, \quad i \leq r \\ a_{ii}x_i &= \sum_{j=i-r}^{i-1} a_{ij}x_j + y_i, \quad i > r \end{aligned} \tag{7.26}$$

I.e, let vector X_a with a 4 by 4 reverse tri-diagonal correlation matrix R_a

$$\mathbf{R_a} = \begin{pmatrix} 1.000 & -0.612 & 0.408 & -0.250 \\ -0.612 & 1.000 & -0.667 & 0.408 \\ 0.408 & -0.667 & 1.000 & -0.612 \\ -0.250 & 0.408 & -0.612 & 1.000 \end{pmatrix}$$

here R_a^{-1} is a tri-diagonal matrix

$$\mathbf{R_a^{-1}} = \begin{pmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 \\ 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{pmatrix}$$

Then X_a is a Craig-AR(1) process. Here we have $X_a = A^{-1}Y$, where A is the decompose matrix of R_a^{-1} that $R_a^{-1} = AA^t$ and Y is k-vector of i.i.d standard norm distribution.

In time series, a process comes from the past and goes to the future. But our process begins in time "1" and end at time "k". The value of x_1 can be considered as initial value which does not depend on the past.

Similarly we could define the Auto Regression Moving Average(ARMR) model. Unfortunately, there is no algorithm available until now to solve ARMR(1,1) probability. So we left these problem as challenging problems of the future.

For simplicity, let X_is have unit variances, zero mean and correlation matrix R_a, which is reverse tridiagonal, the edge orthoscheme probability $P_k(U, R_a)$ can be calculate from a Markov chain

$$\begin{aligned} X_i &\sim N(0,1), & i &= 1 \\ X_{i+1}|X_i &\sim N(\rho_i X_i, \sigma_i^2), & i &> 1 \end{aligned} \tag{7.27}$$

here $\rho_i = corr(X_{i+1}, X_i)$ and $\sigma_i^2 = 1 - \rho_i^2$. Note $R_m = \{\rho_{ij}\}, 1 \le i, j \le m$ as the correlation matrix for random vector $X_{1,m} = \{X_1, ..., X_m\}$. R_m is just sub matrix of R_a and $R_k = R_a$. We can calculate the one-side probability of the edge orthoscheme

by

$$
\begin{aligned}
&P_k(U, R_a) \\
&= \int_{-u_k}^{\infty} ... \int_{-u_1}^{\infty} \phi_k(X; 0, R_a)\, dx_1 ... dx_k \\
&= \int_{-u_k}^{\infty} ... \int_{-u_1}^{\infty} P(x_k|x_{k-1}) \phi_{k-1}(X_{1,k-1}; 0, R_{k-1})\, dx_1 ... dx_{k-1} dx_k \\
&= \int_{-u_k}^{\infty} \int_{-u_{k-1}}^{\infty} P(x_k|x_{k-1}) \int_{-u_{k-2}}^{\infty} ... \int_{-u_1}^{\infty} \phi_{k-1}(X_{1,k-1}; \mathbf{0}, R_{k-1})\, dx_1 ... dx_{k-2} dx_{k-1} dx_k \\
&= \frac{1}{\sigma_k} \int_{-u_k}^{\infty} \int_{-u_{k-1}}^{\infty} \phi_1(\frac{x_k - \rho_{k-1} x_{k-1}}{\sigma_{k-1}}) \int_{-u_{k-2}}^{\infty} ... \int_{-u_1}^{\infty} \phi_{k-1}(X_{1,k-1}; 0, R_{k-1})\, dx_1 ... dx_{k-2} dx_{k-1} dx_k
\end{aligned}
\tag{7.28}
$$

Craig (2008) introduced a recursive computational approach which can calculate this integral in linear time. He defined

$$
\begin{aligned}
f_1(x_1) &= \phi_1(x_1) = \frac{1}{(2\pi)^{1/2}} \exp^{-\frac{x_1^2}{2}} \\
f_i(x_i) &= \frac{1}{\sigma_i} \int_{-u_{i-1}}^{\infty} \phi_1(\frac{x_i - \rho_{i-1} x_{i-1}}{\sigma_{i-1}}) f_{i-1}(x_{i-1}) dx_{i-1}
\end{aligned}
\tag{7.29}
$$

so that the required probability is transformed into

$$
P_k(U, R_a) = \int_{-u_k}^{\infty} f_k(x_k)\, dx_k \tag{7.30}
$$

Then $f_i(y)$s are calculated one by one from fast Fourier transformation(FFT).

Similarly as Miwa's orthoscheme order theorem, Craig also showed that any k-variate orthant probability can be calculated from at most $(k-1)!$ edge orthoscheme probabilities. Furthermore, the total number of edge orthoscheme probabilities needed to be calculated is usually smaller than or equal to the number of edge orthoscheme probabilities.

7.1.4 Examples for calculating mvn

A new *algorithm* argument to **pmvnorm** and **qmvnorm** has been introduced in **mvtnorm** version .9-0 in order to switch between two algorithms: *GenzBretz()* is the default and triggers the use of the above mentioned quasi-randomized Monte-Carlo procedure by Genz and Bretz (1999). Alternatively, *algorithm = Miwa()* applies the procedure discussed here. Both functions can be used to specify hyper-parameters of the algorithm. For *Miwa()*, the argument *steps* defines the number of grid points G to be evaluated.

The following example shows how to calculate the probability

$$\begin{aligned}&\Phi_3(E, D)\\&= \{-1 < x_1 < 1, -4 < x_2 < 4, -2 < x_3 < 2\}.\end{aligned}$$

with mean $U = (0, 0, 0)^t$ and correlation matrix

$$R_3 = \begin{pmatrix} 1 & 1/4 & 1/5 \\ 1/4 & 1 & 1/3 \\ 1/5 & 1/3 & 1 \end{pmatrix}$$

by using the following R code:

"

```
library("mvtnorm")
m <- 3
S <- diag(m)
S[2, 1] <- S[1, 2] <- 1 / 4
S[3, 1] <- S[3, 1] <- 1 / 5
S[3, 2] <- S[3, 2] <- 1 / 3
pmvnorm(lower = -c(1,4,2), upper = c(1,4,2), mean=rep(0, m), sigma = S, algorithm = Miwa())
```

"

The upper limit and lower limit of the integral region are passed by the vectors

Algorithm	$m = 5$		$m = 10$	
	$\rho = \frac{1}{2}$	$\rho = -\frac{1}{2}$	$\rho = \frac{1}{2}$	$\rho = -\frac{1}{2}$
Genz & Bretz ($\varepsilon = 10^{-4}$)	0.08468833	0.001385620	0.008863600	2.376316×10^{-8}
Genz & Bretz ($\varepsilon = 10^{-5}$)	0.08472561	**0.001390769**	0.008863877	**2.319286** $\times 10^{-8}$
Genz & Bretz ($\varepsilon = 10^{-6}$)	**0.08472682**	0.001388424	**0.008862195**	2.671923×10^{-8}
Miwa ($G = 128$)	0.08472222	0.001388889	0.008863235	2.505205×10^{-8}
Exact.	0.08472222	0.001388889	0.008863236	2.505211×10^{-8}

Table 7.1: Value of probabilities with tri-diagonal correlation coefficients, $\rho_{i,i\pm1} = \rho, 1 \leq i \leq m$ and $\rho_{j,i} = 0, \forall |i - j| > 1$. $\rho = 2^{-1}$ or $\rho = -2^{-1}$.

upper and lower. The mean vector and correlation matrix are given by the vector mean and the matrix corr. From the result, we know that $p = 0.6536804$ with given correlation matrix .

7.2 Accuracy and time consumption

In this section, we compare the accuracy and time consumption of the implementation of the algorithm of Miwa et al. (2003) with the default procedure for approximating multivariate normal probabilities in mvtnorm by Genz and Bretz (1999). The experiments were performed using an Intel® Pentium® processor with 2.8 GHz.

Algorithm	m=5		m=9	
	$\rho = \frac{1}{2}$	sec.	$\rho = \frac{1}{2}$	sec.
Genz & Bretz ($\varepsilon = 10^{-4}$)	0.1666398	0.029	0.09998728	0.231
Genz & Bretz ($\varepsilon = 10^{-5}$)	0.1666719	0.132	0.09998277	0.403
Genz & Bretz ($\varepsilon = 10^{-6}$)	0.1666686	0.133	0.09999726	0.431
Miwa ($G = 128$)	0.1666667	0.021	0.09999995	89.921
Exact.	0.1666667		0.10000000	

Table 7.2: Accuracy and time consumption of centered orthant probabilities with correlation coefficients, $\rho_{j,i} = 2^{-1}, i \neq j$, $1 \leq i \leq m$.

7.2.1 Probabilities with tri-diagonal correlation matrix

The exact value of $P_m(,)$ is known if has some special structure. For example, when is a m-dimensional tri-diagonal correlation matrix with correlation coefficients

$$\rho_{j,i} = \begin{cases} -2^{-1} & j = i \pm 1 \\ 0 & |i-j| > 1 \end{cases} \quad 1 \le i \le m$$

the value of $P_m(\mathbf{0},)$ is $((1+m)!)^{-1}$ (Miwa et al., 2003). The accuracy of Miwa algorithm ($G = 128$ grid points) and the Genz & Bretz algorithm (with absolute error tolerance $\varepsilon = 10^{-4}, 10^{-5}, 10^{-6}$) for probabilities with tri-diagonal correlation matrix are compared in Table 7.1.4. In each calculation, we have results with small variance. The values, which do not hold the tolerance error, are marked with bold characters and are underlined in the tables. When the dimension is larger than five, Genz & Bretz' algorithm with error tolerance smaller than 10^{-5} is hard to achieve. While Miwa's algorithm with grid points $G = 128$ achieves error tolerance smaller than 10^{-7}.

Both algorithms are linear in this simplest case and very fast (<0.01 second), so the time consumption is not discussed here.

7.2.2 Centered orthant probabilities

When is the correlation matrix with

$$\rho_{j,i} = 2^{-1},\ i \neq j,\ 1 \le i \le m$$

the value of $P_m(\mathbf{0},)$ is known to be $(1+m)^{-1}$ (Miwa et al., 2003). Accuracy and time consumption of Miwa's algorithm and Genz & Bretz' algorithm for this situation are compared in Table 7.2. As a numerical algorithm, Miwa's algorithm still

Dimension	Miwa ($G = 128$)		Genz & Bretz ($\varepsilon = 10^{-4}$)	
	One-sided	Two-sided	One-sided	Two-sided
$m = 5$	0.021	0.441	0.029	0.085
$m = 6$	0.089	8.731	0.089	0.149
$m = 7$	0.599	156.01	0.083	0.255
$m = 8$	9.956	4hours	0.138	0.233
$m = 9$	89.921	-	0.231	0.392

Table 7.3: Time consumption of centered orthant probabilities (measured in seconds).

has better tolerance error. However, the time consumption of Miwa's algorithm increases none-linearly when the dimension of the orthant probabilities increases. A detail time consumption analysis for both methods is given in Table 7.3. Miwa's algorithm is much slower than Genz & Bretz' algorithm in calculating two-sided orthant probability when the dimension m is larger than 7.

We have implemented an interface to the procedure of Miwa et al. (2003) in the package mvtnorm. For small dimensions, it is an alternative to quasi-randomized Monte-Carlo procedures, which are computed by default. However, Miwa's algorithm has some disadvantages. When the dimension m increases, the time consumption of Miwa's algorithm increases dramatically. Moreover, it can't be applied to singular problems which are common in multiple testing problems, for example.

7.3 Package Binotrend

Package Binotrend is used for testing linear trend in binomial data. Function "binoint()", "Likelihood()" and "Likelihoodep()" are three major functions. "binoint()" runs the MCT developed by Bretz (1999); Bretz and Hothorn (2003). "Likelihood()" uses the IC method developed in these thesis to detect the linear trend. "Likelihoodep()" is a special and fast single function for detecting motif under Epidemic-order.

binoint()

binoint(present,samplesize,type,cmatrix,alternative,conf.level=.95)

here argument *present* and *samplesize* are the number of observations and sample sizes. *type* defines the contrasts type. User can also use *cmatrix* to define their own type. *alternative* can be chosen as one-sided test(*alternative*=less, *alternative*=greater) or two-sided test(two.sided). *conf.level* controls the FWER.

Sample size calculation for MCT

binoint(Pi,type,alternative,expect.power)

here argument *Pi* is the proportion from data. *expect.power* is equal .80 by default.

Likelihood()

Likelihood(X,N,formoflikelihood,penalty)

here argument X and N are the number of observations and sample sizes. *formoflikelihood* defines the IC type. The user can also change *penalty* to have extra bias adjustment. The value of *penalty* is 1.5 by default.

Likelihoodep()

Likelihoodep(X,N,formoflikelihood,penalty)

This function is a special and fast version for doing DNA-motif finding. The value of *penalty* is 2 by default.

Cochran-Armitage Test

CA.test2<-function(success, failure, scores, alternative="two.sided")

This function is written by Schaarschmidt et al. (2009) to make the Cochran-Amitage Test. *success* is the number of present and *failure* is the number of absent. *scores* is the scores among different doses.

7.4 Code for summary section

7.4.1 Single Change-point order restriction

Adverse events rate

```
# data input
X=c(9, 19, 24)
N=c(20,43, 41)


# MCT
binoint(present=X,samplesize=N,type="Changepoint",alternative="less",conf.level=.95)


# Sample size calculation of MCT with given power
binoint(Pi=X/N,type="Changepoint",alternative="two.sided",expect.power=.80)


# MLT
Likelihood(X,N,"changepoint-alpha-control")


# ORIC-lMLE
```

```
Likelihood(X,N,"Anraku")


# AIC
Likelihood(X,N,"Anraku",penalty=2)
```

Cochran-Amitage Test

```
# data input


adverserate <- data.frame(dose = rep(c(0, 1, 2), c(20, 43, 41)),
tumor = c(rep(c(0, 1), c(11, 9)),
rep(c(0, 1), c(24, 19)),
rep(c(0, 1), c(17, 24))))


Success= table(adverserate)[,2]
Failure= table(adverserate)[,1]

# CA test


CA.test2(Success,Failure,score=c(0,1,2))
```

7.4.2 Epidemic-order restriction

DNA-motif finding

```
# data input
```

```
entrmatrix<-matrix(c(0,0,0,7,1,1,9,0,6,1,0,3,1,2,0,0,0,
14,0,1,3,3,6,1,8,0,5,3,7,5,3,12,14,0,
0,14,13,4,9,6,3,5,0,6,1,2,6,1,2,0,
14,0,0,0,0,1,1,1,1,8,2,10,2,2,8,0,0,0),
nrow = 4, ncol=17, byrow=TRUE,dimnames = list(c("A", "C","G","T") ,c(1:17)))

X=14*maxmatrix(entrmatrix)
N=rep(14,17)

# MCT
binoint(present=X,samplesize=N,type="newone",alternative="less",conf.level=.95)

# NIC
Likelihoodep(X,N,"Nino")
```

7.4.3 Simple order restriction

Spontaneous abortion rate

```
# data input
X=c(33, 37, 3, 7)
N=c(259, 358, 64, 12)

# contrasts matrix input

cmatrix=t(matrix(
c(
-3,1,1,1,
```

```
-1,-1,1,1,
-1,-1,-1,3,
-3,-1,1,3,
-1,-1,0,2,
-1,0,0,1,
-2,0,1,1 ),4,7
)
)

# MCT
output=binoint(present=X,samplesize=N,cmatrix=cmatrix,alternative="less",conf.level=.95)

# the output here is very long, so it could be better to use output $ to look into the
detail.

# MLT
Likelihood(X,N,"Isotonic-4-dim-alpha-control")

# MHIC
Likelihood(X,N,"Mi-Hothorn-IC")

# ORIC-lMLE
Likelihood(X,N,"Isotonic-4-dim-Anraku")
```

Cochran-Amitage Test

```
# data input

adverserate <- data.frame(dose = rep(c(0, 1, 2,3), c(259, 358, 64, 12)),
```

```
tumor = c(rep(c(0, 1), c(33, 226)),
rep(c(0, 1), c(37, 321)),
rep(c(0, 1), c(3, 61)),
rep(c(0, 1), c(7, 5))))


Success= table(adverserate)[,2]
Failure= table(adverserate)[,1]

# CA test


CA.test2(Success,Failure,score=c(0,1,2,3))
```

7.4.4 Simple-tree restriction

Adverse events rate

```
# data input
X=c(9, 19, 24)
N=c(20,43, 41)


# MCT

typeofbio="Dunnett"
binoint(present=X,samplesize=N,type=typeofbio,alternative="less")
```

Different from the result given before, function *Likelihood()* returns log-likelihood ratios directly. We also use *PenaltySimpleTreeAlphaControl()* calculate the critical value for MLT method

```
# MLT
```

```
CV=PenaltySimpleTreeAlphaControl(k=3,x=X,n=N,flag=1)
Likelihood(X,N,"log-ratio-simple-tree",penalty=CV-1)

# ORIC-lMLE
Likelihood(X,N,"log-ratio-simple-tree",penalty=1.5-1)
```

Chapter 8

Summary

In this chapter, we will give a solution to the previous examples. Four methods, which are MLT, MCT, ORIC-lMLE and MHIC, will be compared together for the power, CR and MR. The final conclusion and open questions will be discussed in the end.

8.1 Solution to the previous examples

8.1.1 Single Change-point order restriction

Adverse events rate

We can calculate the log-likelihood and the ICs. MCT and MLT are also listed in for comparison. Under Single Change-point order restriction, all these methods will use the MLE as the best estimator.

Method	$H_0(T_{max})$	H_A^1	H_A^2	α	$1-\beta$	selected model
MCT	(1.898)	0.511	1.349	.050	.317	H_0
log-likelihood	-6.902	-6.779	-5.913			
MLT	-7.902	-9.624	-8.759	.050	.317	H_0
MHIC	-7.902	-8.279	-7.413	N.	.720	H_A^2
ORIC-lMLE	-7.902	-8.279	-7.413	N.	.720	H_A^2
AIC	-7.902	-8.779	-7.913	N.	N.	H_A^2

here the penalty term are given as following

Penalty	H_0	H_A^1	H_A^2
MLT	1	2.95	2.95
MHIC	1	1.5	1.5
ORIC-lMLE	1	1.5	1.5
AIC	1	2	2

Comparisons with Cochran-Armitage Test

In the following equations, we will calculate LRT and CAT from our last example.

Method	$H_0(T_{max})$	H_A^1	H_A^2	chi-square(ratio)	DF	p-value
log-likelihood	-6.902	-6.779	-5.913	(0.989)	-	
LRT	-	-	-	(1.878)	1	0.1705
CAT'	-	-		1.422	1	0.1161(one-sided)

CAT has a smaller p-value in this case. It is more sensitive to detect the linear trend. However it is only a trend test, which does not assume any alternative model. So, it cannot do model selection and has no indication of the back ground structures.

Results

From the result we can see that AIC, MHIC and ORIC-lMLE select model H_A^2 and draw the conclusion that higher dose of such drug has the adverse effect. Since these three methods are model selection methods, none of them can test the null hypothesis

	s=13	s=14	s=15	s=16
r=1	-51.2	-44.9	-50.5	-57.1
r=2	-45.4	-38.1	-43.3	-50.5
r=3	-40.6	*-33.4*	-35.1	-42.3
r=4	-46.7	-42.7	-45.4	-51.2

Table 8.1: Adjusted log-likelihood value of the DNA problem

with certain α level. On the other hand, MCT and MLT test the hypothesis. They cannot reject the null hypothesis with certain FWER level, but they calculate the power and suggest increase the sample to get a higher power. For example, in order to achieve higher power of 80% i.e. $1 - \beta = .80$, we need the total sample size increase to 417 for blanched case (Bretz et al., 2005). After that they will probably select model H_A^2 as the best model.

8.1.2 Epidemic-order restriction

NIC method

The MLEs are calculated similarly as Single Change-point order restriction. The log-likelihood is calculated with given MLE and the penalty term is one under null hypothesis and under alternative hypothesis. The later penalty term is the sum of penalty from two Change-points and two unknown parameters (Ninomiya, 2005). Since we have two unknown Change-points, the result are listed in a two dimensions table. Part of the table is given in Table 8.1 and the value of the null model is -54.1. We can select position "3" and "14" as the best prediction of the Change-points.

MCT method

We can also use MCT to find the Change-points. The results are shown in following pictures. The pattern in the circle is the possible pattern ranked from 1 to 5. With comparing these five patterns, the most possible pattern and its "neighbors" are founded. The asymptotic power can also be calculated.

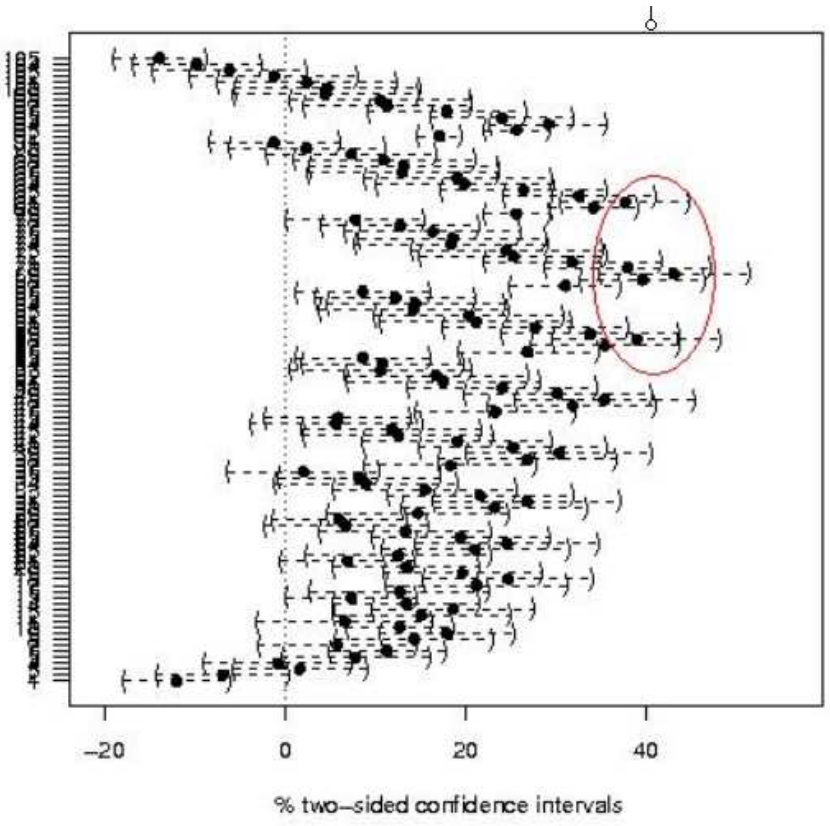

Figure 8.1: Simultaneous confidence intervals for all possible models. Here we plot the value test statistics (the black points) and their intervals for different models simultaneously. The largest value is obtained by model $H_A^{r=3,s=13}$. We also find that model $H_A^{r=2,s=13}$, $H_A^{r=3,s=14}$, $H_A^{r=3,s=12}$ and $H_A^{r=3,s=14}$ also have relatively larger value among others.

Rank 1 : 11 11 11 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 11 11 11
Rank 2: 12 12 12 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 12 12
.....
Rank 5 : 12 12 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 12 12 12

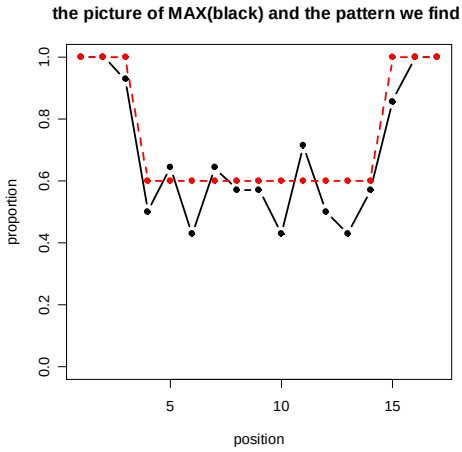

Figure 8.2: Contrasts for the top 5 pattern and the entropy comparison of the most possible pattern.

Meth.	$H_0(T_{max})$	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5	H_A^6	H_A^7	α	$1-\beta$	Sel.
MCT	(1.83)	2.97	3.84	5.24	4.57	5.02	4.83	3.47	.05	.94	H_A^3
MLT	-19.08	-20.76	-20.66	-12.96	-22.92	-20.39	-24.91	-27.08	.05	.94	H_A^3
MHIC	-19.08	-19.58	-19.48	-11.78	-21.74	-19.21	-23.73	-25.90	N	.97	H_A^3
ORIC-lMLE	-19.08	-19.58	-19.48	-11.78	-12.35	-12.11	-12.11	-19.81	N	N	H_A^3

Table 8.2: Value of the ICs

Comparision with Cochran-Armitage Test

CAT is a trend test which requires monotone trend. The expected structure here is concave. CAT is not available for this problem.

Results

From the result we can see that NIC and MCT select model $H_A^{3,14}$ and detect two Change-points at position "3" and "14". NIC cannot reject the null hypothesis with certain α level. But from the simulation study we know that it has very good error control, While MCT controls FWER level. But FWER is not the topic here. The researchers care about "power" more than "Error control". Also the calculation time for multivariate normal distribution is very long if we have hundreds of possible patterns. So IC method is the simple and fast method we recommend. By using NIC, we find pattern "CGG—GCG" as the most possible pattern of motif. We also suggest check pattern "CG—GCG" and CGG—CG".

8.1.3 Simple-order restriction

Spontaneous abortion rate

The final ICs are given in Table 8.2

The the penalties are liested in Table 8.3

Penalty	H_0	H_A^1	H_A^2	H_A^3	H_A^4	H_A^5	H_A^6	H_A^7
MLT	1	3.23	3.23	3.23	3.23	3.23	3.23	3.23
MHIC	1	1.5	1.5	1.5	1.5	1.5	1.5	1.5
ORIC-lMLE	1	1.5	1.5	1.5	2.07	1.83	1.83	1.83

Table 8.3: Penalties of the ICs

Comparisons with Cochran-Armitage Test

We get the test statistics of LRT and CAT. In the following equations, we will calculate them from our last example.

Method	$H_0(T_{max})$	H_A^3	chi-square(ratio)	DF	p-value
log-likelihood	-19.089	-11.781	(7.308)	-	
LRT	-	-	(14.616)	3	0.002
CAT'	-		0.4167	1	0.2093(one-sided)

LRT has a smaller p-value in this case. It also rejects the null hypothesis. CAT fails to reject the null hypothesis. By comparing the result from Section 8.1.1, we know that CAT is not always powerful. It is very powerful, if there is a strong linear trend in the data structure.

Results

All methods detect a trend and select model H_j^3 as the best model. Because both MHIC and ORIC-lMLE methods are model selection methods, we cannot control the FWER by using these two methods. For MLT and MCT methods, we select model H_j^3 as the best model with FWER control. The asymptotic power is 0.94.

The conclusion for the problem is that we select model H_j^3 as the best model, i.e. when the age of the father is higher than 35, the mother's abortion rate will be much higher than others.

8.1.4 Simple-tree order restriction

Adverse events rate

Finally, we can calculate the log-likelihood *ratio* and the ICs. MCT and MLT are also listed in for comparison. Under Simple-tree order restriction, all these methods will use the MLE as the best estimator.

Method	$H_0(T_{max})$	H_A^1	H_A^2	α	$1-\beta$	selected model
MCT	(1.882)	-0.060	0.992	0.050	0.193	H_0
MLT-ratio	0.000	-1.771	-1.276	0.050	0.193	H_0
ORIC-lMLE	0	-0.500	-0.004	N.	-	H_0

here the penalty term are given as following

Penalty	H_0	H_A^1	H_A^2
MLT	1	2.77	2.77
MHIC	1	1.5	1.5
ORIC-lMLE	1	1.5	1.5

Comparisons with Cochran-Armitage Test

We are testing many dose to the control group. there is no trend here. CAT is not available for this many-to-one comparison.

Results

The test for Simple-tree order tests 2 groups at one time for each alternative model. The sample size here is so small that we cannot reject the null hypothesis. No dose level is significantly different to the control group.

In order to achieve higher power of 80% i.e. $1-\beta = .80$, we need the total sample size increase to 612 for blanched case (Bretz et al., 2005). After that they will probably select mode H_A^2 as the best model.

8.2 Conclusions

8.2.1 Main results

We have developed Multiple Log-likelihood Test(MLT) with FWER controlled under different order restrictions. For MLT method, the penalty term is also the critical value of a test, so *it selects the model and controls the α rate at the same time. We can consider it as a bridge between contrasts method and IC method.*

$$2 * (MLT(H_A^j) - MLT(H_0)) \geq MCT(H_A^j)^2 \tag{8.1}$$

MLT, which is identical to MCT for normal and binomial data in order restricted problem, with lower α ($\alpha \approx 25\%$) rate, can be used for model selection for 3 types: Many-to-one (Simple-tree), Change-point and Simple-order.

MCT defines different contrasts for all elementary alternative models and tests all of them in a multiple contrast test; after rejecting the null for at least one alternatives, *selects the one with the largest test statistics.* While, AIC, MHIC, ORIC-lMLE and MLT method use Information Criterion and *select the models with the largest adjusted log-likelihood.*

8.2.2 The relationship

The relationships among MCT, MLT and ICs can be described in detail as following:

MLT and MCT

The ICs from MLT and the test statistics of MCT satisfy equation (8.1) under the alternative model.

Under Single Change-point order restriction, the condition in Equation 5.46 is sat-

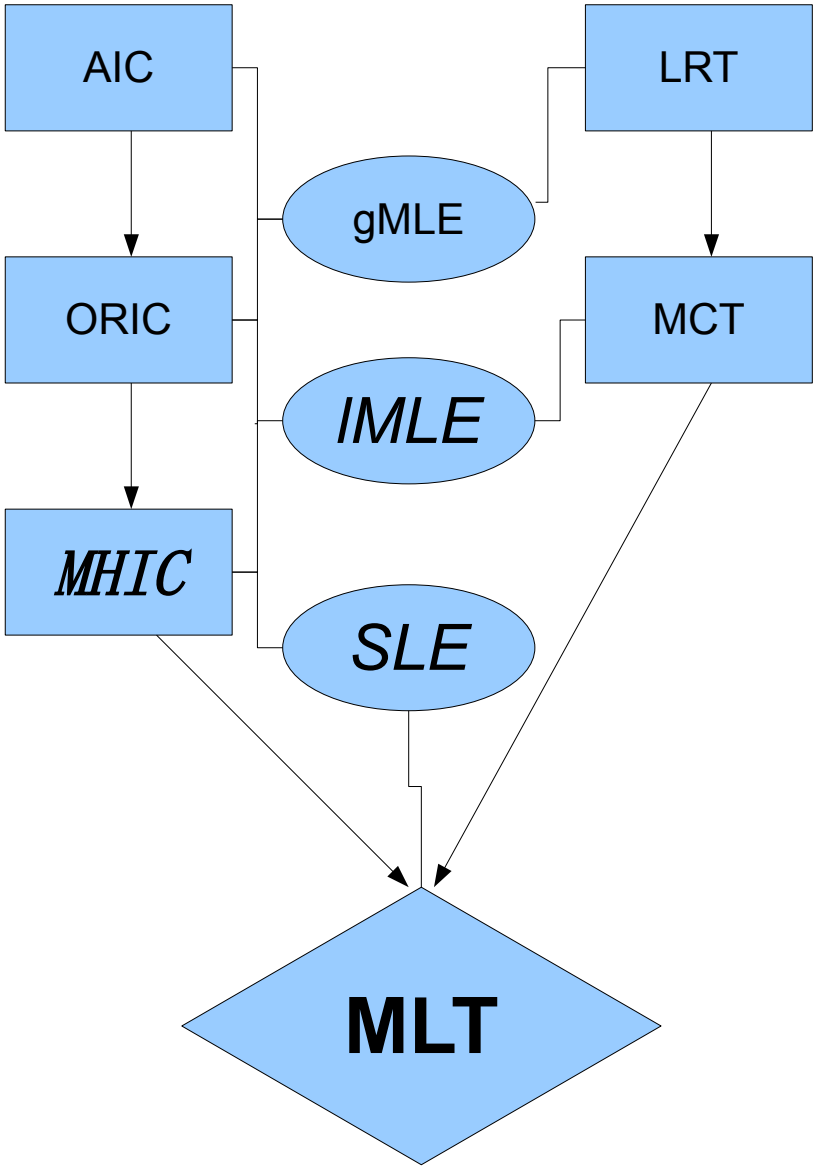

Figure 8.3: The relationship of test-based method, model-based method and our new method. The new creations are marked in bold

isfied and the equality holds

$$2 * (MLT(H_A^j) - MLT(H_0)) = MCT(H_A^j)^2 \tag{8.2}$$

The critical value of MLT using SLE, can be calculated from chi-square distribution

$$z_{1-\alpha} = 0.5 Z_{q,1-\alpha}^2 \tag{8.3}$$

$Z_{q,1-\alpha}$ is the α quantile for q-variate normal distribution.

MLT and ICs

For given null model the ICs of AIC, ORIC-lMLE MLT and MHIC have the following relationship

$$MLT(H_0) = MHIC(H_0) = AIC(H_0) = ORIC - lMLE(H_0) \tag{8.4}$$

For given elementary alternative model the ICs of MLT and MHIC have the following relationship

$$ORIC - lMLE(H_A^j) \geq MHIC(H_A^j) = MLT(H_A^j) - 1.5 + z_{1-\alpha} \tag{8.5}$$

For given elementary alternative model under Single Change-point order restriction, the ICs of MLT and MHIC have the following relationship

$$AIC(H_A^j) + 0.5 = ORIC - lMLE(H_A^j) = MHIC(H_A^j) = MLT(H_A^j) - 1.5 + z_{1-\alpha} \tag{8.6}$$

The Bridge

Using the relationship given above we have

$$\begin{aligned} & 2 * ((ORIC - lMLE(H_A^j) + 1.5 - z_{1-\alpha}) - ORIC - lMLE(H_0)) \\ \geq & 2 * (MLT(H_A^j) - MLT(H_0)) \\ \geq & MCT(H_A^j)^2 \end{aligned} \tag{8.7}$$

Under Single Change-point order restriction, the condition in Equation 5.46 is fulfilled and the equality holds.

Outlook

This bridge can be extended to any MCT that uses ordered contrast. For any MCT, we can use the same SLE to build a Multiple Likelihood Test, which is test-based model selection procedure and has similar power behavior as the correspond MCT. Furthermore, we can use the same SLE to build Mi and Hothorn Information Criterion, which has higher correct model selection rate than others, to do model selection.

For further development of this extension and power estimation, we need a powerful software which can calculate weighted multivariate chi-square distribution.

For small sample size problems, the variance estimators and confidence intervals could have be improved by using Add-4-method (Schaarschmidt et al., 2009).

Bibliography

Abelson, R. P. and Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative-analysis - general-theory and case of simple order. *Annals Of Mathematical Statistics*, 34(4):1347–&.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, New York.

Agresti, A. and Caffo, B. (2000). Simple and effective confidence interval for proportions and differences of proportions result from adding two successes and two failures. *Am.Stat*, 54:280–288.

Agresti, A. and Coull, B. A. (1996). Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics*, 52(3):1103–1111.

Akaike, H. (1974). New look at statistical-model identification. *Ieee Transactions On Automatic Control*, AC19(6):716–723.

Akaike, H. and Kitagawa, G. (1998). *The Practise of Time Series Analysis*. Springer.

Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika*, 86(1):141–152.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386.

Bartholomew, D. J. (1959). A test of homogeneity for ordered alternatives. *Biometrika*, 46(3-4):328–335.

Bretz, F. (1999). *Powerful Modifications of Williams Test on Trend*. PhD thesis, University Hannover.

Bretz, F. and Hothorn, L. A. (2002). Detecting dose-response using contrasts: asymptotic power and sample size determination for binomial data. *Statistics in Medicine*, 21(22):3325–3335.

Bretz, F. and Hothorn, L. A. (2003). Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. *Atla-Alternatives to Laboratory Animals*, 31:81–96.

Bretz, F., Pinheiro, J. C., and Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61(3):738–748.

Burnham, K. P. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical-Theoretic Approach, 2nd ed.* Springer.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference - understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304.

Chaudhuri, S. and Perlman, M. D. (2005). On the bias and mean-square error of order-restricted maximum likelihood estimators. *Journal of Statistical Planning and Inference*, 130(1-2):229–250.

Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometrics*, 10(4):417–451.

Craig, P. (2008). A new reconstruction of multivariate normal orthant probabilities. *Journal of The Royal Statistical Society Series B-Statistical Methodology*, 70:227–243.

Dosemeci, M. and Benichou, J. (1998). An alternative test for trend in exposure-response analysis. *Journal of Exposure Analysis and Environmental Epidemiology*, 8(1):9–15.

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.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–149.

Genz, A. and Bretz, F. (1999). Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63:361–378.

Grimsrud, T. K., Berge, S. R., Haldorsen, T., and Andersen, A. (2002). Exposure to different forms of nickel and risk of lung cancer. *American Journal Of Epidemiology*, 156(12):1123–1132.

Halpern, A. L. (1999). Minimally selected p and other tests for a single abrupt changepoint in a binary sequence. *Biometrics*, 55(4):1044–1050.

Hirotsu, C. and Marumo, K. (2002). Changepoint analysis as a method for isotonic inference. *Scandinavian Journal of Statistics*, 29(1):125–138.

Hirotsu, C. and Srivastava, M. S. (2000). Simultaneous confidence intervals based on one-sided max t test. *Statistics & Probability Letters*, 49(1):25–37.

Hothorn, L. A., Vaeth, M., and Hothorn, T. (2008). *Trend tests for the evaluation of exposure-response relationships in epidemiological exposure studies*. Epidemiologic Perspectives & Innovations.

Hothorn, T., Bretz, F., and Genz, A. (2001). On multivariate t and Gauß probabilities in R. *R News*, 1(2):27–29.

Hughes, A. W. and King, M. L. (2003). Model selection using AIC in the presence of one-sided information. *Journal of Statistical Planning and Inference*, 115(2):397–411.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer.

Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins-Structure Function And Genetics*, 7(1):41–51.

Lewin, B. (2004). *Genes VIII.* Prentice Hall International.

McDermott, M. P. (1999). Generalized orthogonal contrast tests for homogeneity of ordered means. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 27(3):457–470.

Miwa, T., Hayter, A. J., and Kuriki, S. (2003). The evaluation of general non-centred orthant probabilities. *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, 65:223–234.

Mukerjee, H., Robertson, T., and Wright, F. T. (1987). Comparison of several treatments with a control using multiple contrasts. *Journal Of The American Statistical Association*, 82(399):902–910.

Ninomiya, Y. (2005). Information criterion for gaussian change-point model. *Statistics & Probability Letters*, 72(3):237–247.

Ninomiya, Y. (2006). Personal comunication.

Pevzner, P. A., Tang, H. X., and Waterman, M. S. (2001). An eulerian path approach to dna fragment assembly. *Proceedings of The National Academy of Sciences of The United States of America*, 98(17):9748–9753.

Pincus, R. (1975). Testing linear hypotheses under restricted alternatives. *Mathematische Operationsforschung und Statistik*, 6:733–751.

Roberts, S. and Martin, M. A. (2006). The question of nonlinearity in the dose-response relation between particulate matter air pollution and mortality: Can akaike's information criterion be trusted to take the right turn? *American Journal of Epidemiology*, 164(12):1242–1250.

Robertson, T., Wright, F., and Dykstra, R. (1988). *Order restricted statistical inference.* Wiley, New York.

Royston, P., Ambler, G., and Sauerbrei, W. (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28(5):964–974.

Schaarschmidt, F., Biesheuvel, E., and Hothorn, L. A. (2009). Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials. *Journal of Biopharmaceutical Statistics*, 19(2):292–310.

Slama, R., Werwatz, A., Boutou, O., Ducot, B., Spira, A., and Hardle, W. (2003). Does male age affect the risk of spontaneous abortion? an approach using semiparametric regression. *American Journal of Epidemiology*, 157(9):815–824.

Srivastava, M. and Carter, E. (1975). *Applied Multivariate Statisticas*. North-Holland.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the perceptron algorithm to distinguish translational initiation sites in escherichia-coli. *Nucleic Acids Research*, 10(9):2997–3011.

van Zwet, E. W., Kechris, K. J., Bickel, P. J., and Eisen, M. B. (2005). Estimating motifs under order restrictions. *Statistical Applications in Genetics and Molecular Biology*, 4:1.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.

Wright, F. T. (1988). The one-way analysis of variance with ordered-alternatives: A modification of *B*artholomew $\bar{E}^2$ test. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 16(1):75–85.

Xiong, C. and El Barmi, H. (2002). On detecting change in likelihood ratio ordering. *Journal of Nonparametric Statistics*, 14(5):555–568.

Yao, Q. W. (1993). Tests for change-points with epidemic alternatives. *Biometrika*, 80(1):179–191.

Zhao, L. C. and Peng, L. M. (2002). Model selection under order restriction. *Statistics & Probability Letters*, 57(4):301–306.

Zhu, J. and Zhang, M. Q. (1999). Scpd: a promoter database of the yeast saccharomyces cerevisiae. *Bioinformatics*, 15(7-8):607–611.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, 44(1):41–61.

Acknowlegement

I would like to thank all that have helped me. With their help, I have completed this.

I thank Prof. Hothorn for his guidance, encouragement and support over the past three years. This thesis is based on our former research. I thank all my colleagues at Biostatistics Unit of Leibniz University Hannover for their contributions and support. I thank Frank Schaarschmidt, Daniel Gerhard, Kornelius Rohmeyer, Mario Hasler, Ralph Scherer and Junjie Fu for discussing problems and reviewing the thesis. Hannelore Visser and Clemens Buczilowski have provided their technical support.

My special thanks goes to my wife, Xiaozhu Mei. Thank her for her patience, support and love. Last but not the least, I thank my parents. I am always indebted to their love.

Xuefei Mi(Chinese, 13.05.1980)

website: www.biostat.uni-hannover.de/staff/mi/ email: mi@pz.uni-hohenheim.de

Research Interest: Multiple inference. Information Criterion and Model Selection, Multi-stage selection in Plants Breeding.

Education

University of Hannover 06.2005 - 06.2008
Doctor in Biostatistics Advisor: Ludwig A. Hothorn
PhD. Thesis : Model Selection Procedure with Familywise Error Rate Control for Binomial Order-Restricted Problems

University of Kaiserslautern 2003 - 2005
Master Degree in Financial Mathematics GPA 1.9/1.0

Research Experience

University of Hohenheim Postdoctoral Fellow(Wiss. Mitarbeiter)
Hohenheim, Germany 07.2008 - ...

Working in Plants Breeding GABI-GAIN workgroup-D as project leader. Developing algorithm and mathematical models by using the gnomic information from Plant Breeding. These studies can be applied in energy and economic crops.

University of Hannover Research assistant(Wiss. Mitarbeiter)
Hannover, Germany 06.2005 - 06.2008

Researching on Multiple Inference, developing algorithm and regression models for model selection, making programs by R and collaborating with the research team to apply and validate the theory of model selection in Clinical trials Phase II, Phase III, Epidemiology and Case-Control Studies.

Major Publications and Conference Talks

New Numerical Algorithm for Multivariate Normal Probabilities in mvtnorm Package
R Journal issue 1, 2009 2009
Authors: Xuefei Mi, Tetsuhisa Miwa and Torsten Hothorn

Model Selection under Change Point Order Restriction
5th International Conference on Multiple Comparison Procedures,Vienna 2007
Authors: Xuefei Mi, Ludwig Hothorn

DNA-Motif Identification Using Multiple Contrasts
23rd International Biometric Conference, Montreal 2006
Authors: Xuefei Mi, Ludwig Hothorn

Awards

Survivability model of Three Gorges
This is supported by funding "Tiao Zhan Bei" (Peking University) 2002
Authors: Shanshan Xu, Xuefei Mi

Awards for Good Leadership in Student Union
By Beijing University of of Chemical Technology 2000

Qualifications

Programming Languages and Tools:C,C++, R,MatLab, Mathematica, Latex, SAS.

Languages: English, German(DSI passed), Chinese(native speaker)

Eidesstattliche Erklärung zur Dissertation

Hierdurch erkläre ich an Eides statt, dass die Dissertation

„Model Selection Procedure with Familywise Error Rate Control for Binomial Order-Restricted Problems “

selbstständig verfasst und alle benutzten Hilfsmittel sowie evtl. zur Hilfeleistung herangezogene Institutionen vollständig angegeben wurden.

Die Dissertation wurde nicht schon als Diplom- oder ähnliche Prüfungsarbeit verwendet.

Hohenheim, den 18.06.09

_____

(Unterschrift)

Name: Xuefei Mi

Xuefei Mi

Research Interest: Multiple Inferences, Information Criterion and Model Selection.

Full list of Publications

A Model Selection procedure with FWER-control for the binomial change point problem
Writing Current
Authors: Xuefei Mi, Ludwig Hothorn

Implement of Miwa's analytical algorithm of multi-normal distribution
R Journal issue 1, 2009
Authors: Xuefei Mi, Miwa

Model Selection under Change Point Order Restriction
5th International Conference on Multiple Comparison Procedures,Vienna July 2007
Authors: Xuefei Mi, Ludwig Hothorn

DNA-Motif Identification Using Multiple Contrasts
23rd International Biometric Conference, Montreal July 2006
Authors: Xuefei Mi, Ludwig Hothorn

An method of choosing suitable transformations and its application
Master thesis 2005
Advisor: Juergen Franke

Automatic recognizing of the texts in internet
Scientific Chinese 2003, 11, 53
Authors: Baoya Chen, Xuefei Mi, etc.

The probability characters in motif finding
Bachelor thesis 2003
Advisor: Yongyu Yang

Survivability model of Three Gorges
This is supported by funding "Tiao Zhan Bei" (Peking University) 2002
Authors: Shanshan Xu, Xuefei Mi