

國立交通大學

統計學研究所

碩 士 論 文

貝氏方法在多選題排序上的應用

Bayesian Ranking Responses in Multiple-Choice Questions



研 究 生：張少源

指導教授：王秀瑛 教授

中 華 民 國 九 十 九 年 六 月

貝氏方法在多選題排序上的應用
Bayesian Ranking Responses in Multiple-Choice Questions


研 究 生：張少源

Student：Shao-Yuan Chang

指導教授：王秀瑛

Advisor：Hsiuying Wang

國 立 交 通 大 學
統 計 學 研 究 所
碩 士 論 文



A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics

June 2010

Hsinchu, Taiwan

中華民國九十九年六月

貝氏方法在多選題排序上的應用

學生：張少源

指導教授：王秀瑛教授

國立交通大學統計學研究所碩士班

摘 要

在許多調查研究中，問卷調查是一個很重要的工具。許多文獻上對於可複選的問題分析不如研究單選問題那麼的深入。Wang (2008a)在 frequentist 的架構下，提出針對複選題作排序的方法。但是在實際的情況下，對於各個選項也許存在著事前分配，所以建立新的方法結合過去資料與新的資料作排序在問卷調查中是必要的課題。在本篇研究中，我們根據貝式多重檢定的方法，藉由控制後驗的錯誤發生率來得到在貝式架構下的排序。除此之外，我們也將用模擬的方法去比較這些方法的差異及恰當的拒絕區域。

Bayesian Ranking Responses in Multiple-Choice Questions

student : Shao-Yuan Chang

Advisors : Prof. Hsiuying Wang

Institute of Statistics
National Chiao Tung University

ABSTRACT

In many studies, the questionnaire is an important tool for surveying. In the literature, the analyses of multiple-choice questions are not established as in depth as those for single-choice question. Wang (2008a) proposed several methods for ranking the Responses in Multiple-Choice Questions under the usual frequentist setup. However in many situations, there may exist prior information for the ranks of the responses, therefore, establishing a methodology combining the update survey data and the past information for ranking the responses is an essential issue for the questionnaire data analysis. In this paper, we based on several Bayesian multiple testing procedures to develop the Bayesian ranking methods by controlling the posterior expected false discovery rate. In addition, a simulation study is conducted to make a comparison of these approaches and to derive the appropriate rejection region for the testing.

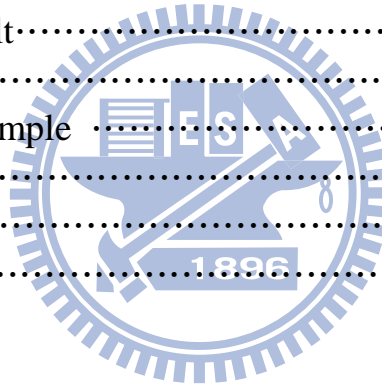
誌 謝

謝謝家人在碩士班這兩年背後的支持與鼓勵，也感謝指導老師對我的關心和論文上的指導與協助，最後謝謝幾個好朋友，我會記得這些日子的。



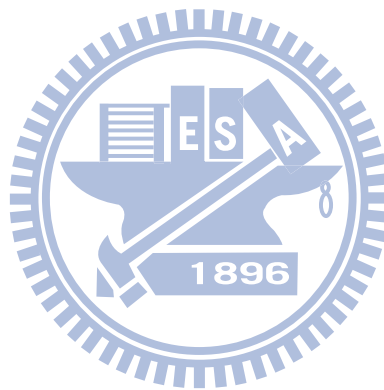
目 錄

中文提要	i
英文提要	ii
誌謝	iii
目錄	iv
表目錄	v
圖目錄	vi
1 、	Introduction.....	1
2 、	Model	4
2.1	Model Selection	5
3 、	Testing Approach	7
3.1	Multiple Testing	7
3.2	Testing Procedures	10
4 、	Ranking Approach and Ranking Consistency.....	11
4.1	Penalty Score	12
5 、	Simulation Result.....	13
5.1	Rejection Rate	13
6 、	A Real Data Example	16
7 、	Conclusion	19
8 、	Appendix	19
Reference	21



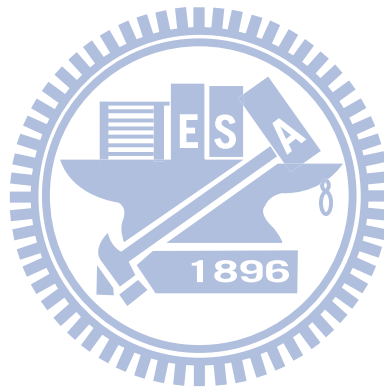
表目錄

Table 1	Outcomes of multiple test.	8
Table 2	The rejection rates for the three methods corresponding to each hypothesis in (4) for 1000 replicates.	14
Table 3	The rejection rates for the three methods corresponding to each hypothesis in (4) for 1000 replicates.	15



圖目錄

Figure 1	The expected penalty scores of the three methods under the condition of Example 1.	24
Figure 2	The expected penalty scores of the three methods under the condition of Example 2.	25



1 Introduction

The questionnaire method is a widely-used tool for researchers in many fields to collect information. It is used especially in marketing or management studies. There are two kinds of questions: single response questions and multiple responses questions. The analyses of multiple responses questions are not as established in depth as those for single response questions. Approaches of analyzing multiple responses questions have been lacking until recently. Umesh (1995) first discussed the problem of analyzing multiple responses questions. Loughin and Scherer (1998), Decady and Thomas (1999) and Bilder, Loughin and Nettleton (2000) propose several methods for testing marginal independence between a single response question and a multiple responses question. Agresti and Liu (1999,2001) discuss the modeling of multiple responses questions. These studies mainly focus on the analysis of the dependence between a single response question and a multiple responses question. However, most researchers are also interested in ranking the responses in a question according to the probabilities of responses being chosen. In fact, the ranking responses problem may be the primary issue in the study of a survey.

Wang (2008a) proposed several approaches to solve this problem. However, these methodologies are derived under the frequestist setup, which cannot be adopted in the Bayesian framework. In real applications, empirical information may exist for the probabilities of responses being chosen. Related applications can refer to Pammer, Fong and Arnold (2000), etc. An appropriate methodology which combines the current data with the past information can provide a more objective ranking strategy than an approach based only on current data. Thus, this study proposes several methods for ranking the responses in a multiple responses question under the Bayesian framework. The methodologies are an extension of the methods proposed in Muller, Parmigiani, Robert and Rousseau (2004). More details about Bayesian multiple testing and applications are discussed by Gopalan and Berry (1998), Do et al. (2005), Gonen, Westfall, Johnson (2003), Scott and

Berger (2006), Muller, Parmigiani and Rice (2007) and Scott (2009).

A related study about Bayesian ranking was carried out by Berger and Deely (2008). Their approach is to rank the items based on the posterior probability of the null hypothesis or Bayes factor. Although the methodology provides a rule for ranking, it does not set up the error tolerance. In the methods used in this study, the statistic used for ranking is similar to the one proposed by Berger and Deely (2008). Furthermore, we also propose the FDR criterion to measure the testing error. In the Bayesian framework, the conventional approach does not associate a criterion to set up the error tolerance. Based on Muller's approach, we can control the testing error within a tolerance level. From this viewpoint, using the Bayesian FDR approach to rank responses is more informative and useful than the conventional approach.

In addition, the Berger and Deelys' approach cannot directly be applied to analyze multiple responses questions. In this study, we clearly illustrate the use of the Bayesian model for analyzing multiple responses questions and derive the exact and approximate Bayes estimator forms. The proposed method can provide a convenient way for researchers to directly adopt the formulas for ranking the responses for multiple responses questions.

First, we use the example described in Wang (2008a) to illustrate the problem. A company is designing a marketing survey to help develop an insect killer. The researchers list several factors, including high quality, price, packaging and smell which could affect the sales market. Thus, the researchers want to know the rank of significance of these factors such that they can design a product with lower cost and higher profit. To obtain the data, a group of individuals are surveyed about purchasing an insect killer. They are asked to fill out questionnaires which list all the questions that addressed to each respondent. The following is the multiple responses question in the questionnaire:

Question 1: Which factors are important to you when considering the purchase of an indoor insect killer ? (1) price (2) high quality (3) packaging (4) smell

(5) others.

In this multiple responses question, there are a total of $2^5 - 1 = 31$ kinds of possible answers because we exclude the case which respondents do not select any response. The 31 random variables constitute a multinomial distribution with multinomial proportions $p \in P = \{p_{i_1 i_2 i_3 i_4 i_5}, i_j = 0 \text{ or } 1 \text{ and } 0 < \sum_{j=1}^5 i_j \leq 5\}$, where i_j cannot be simultaneously equal to 0. Note that the requirement of a multiple responses question is that at least one response is selected. This is not equivalent to a true-false question with five items. If we allow respondents not to select any item or to select all items, it would be equivalent to the five true-false items question. The method developed in this study can extend to this situation.

If we consider the parameter space under the frequentist framework instead of the Bayesian framework. Wang (2008a) provides examples showing that the conventional testing approaches do not possess the property of ranking consistency. This property is a reasonable criterion to reflect the validity of the testing approach. Under the frequentist framework, it is still unknown if a satisfactory approach exists to ranking responses with the property of ranking consistency. In this study, in addition to proposing a ranking approach under the Bayesian framework, a Bayesian ranking consistency property is introduced and the proposed method is shown to be Bayesian ranking consistent.

In the Bayesian framework, assume that we have prior information on the parameter space P and we rank the responses based on a survey study under this prior information. This problem is related to the usual Bayesian multiple testing problem if we consider a single response question. However, the application is more complicated when analyzing multiple responses questions. Muller et al. (2004) proposed several criteria for the Bayesian multiple testing. Miranda-Moreno, Labbe and Fu (2007) applied the methods to hotspot identification in an engineering study. Wang (2008b) carried out a related study estimating the proportions in a multinomial distribution. In this paper, we investigate these Bayesian multiple testing procedures and extend the approaches to rank the responses for multiple

responses questions.

The paper is organized as follows. In Section 2, we describe a Bayesian model for multiple responses responses. Section 3 proposes several Bayesian multiple testing procedures for testing an order of the responses are proposed in Section 3. In Section 4, a ranking criterion is proposed to rank the responses. In addition, the Bayesian multiple testing procedures discussed in Section 3 are shown to be consistent. In Section 5, we present simulation studies to compare the rejection rates of the methodologies and appropriate false discovery rate tolerances for different testing procedures. Finally, Section 6 provides a data example which is ranking inconsistent under the frequentist framework, but is ranking consistent under the Bayesian framework. Finally, a conclusion is given in Section 7.

2 Model

For the general case, assume that a multiple responses question has k responses, v_1, \dots, v_k , and we interview n respondents. Each respondent is asked to choose at least one and at most s answers for this question, where $0 < s \leq k$. If $s = 1$, it is a single response question. There are a total of $c = C_1^k + \dots + C_s^k$ possible kinds of answers that respondents can choose. Let $n_{i_1 \dots i_k}$ denote the number of respondents selecting the responses v_h and not selecting $v_{h'}$ if $i_h = 1$ and $i_{h'} = 0$, and $p_{i_1 \dots i_k}$ denotes the corresponding probability. For example, when $k = 7$, $n_{0100100}$ denote the number of respondents selecting the second and the fifth responses and not selecting the other responses. Thus, the pmf function of $n^* = \{n_{i_1 \dots i_k}, i_j = 0 \text{ or } 1 \text{ and } 0 < \sum_{j=1}^k i_j \leq s\}$ is

$$f_s(n^*) = I(0 < \sum_{j=1}^k i_j \leq s) \frac{n!}{\prod_{i_j=0 \text{ or } 1} n_{i_1 \dots i_k}!} \prod_{i_j=0 \text{ or } 1} p_{i_1 \dots i_k}^{n_{i_1 \dots i_k}}, \quad (1)$$

where $I(\cdot)$ denotes the indicator function. Let m_j denote the sum of the number $n_{i_1 \dots i_k}$ such that the j th response is selected, and π_j denote the corresponding probability, that is $m_j = \sum_{i_j=1} n_{i_1 \dots i_k}$ and $\pi_j = \sum_{i_j=1} p_{i_1 \dots i_k}$. Note π_j is called a marginal

probability of response j . Also let m_{jl} denote the sum of the number n_{i_1, \dots, i_k} such that the j th and l th responses are selected, and π_{jl} denote the corresponding probability. Then $m_{jl} = \sum_{i_j=i_l=1} n_{i_1, \dots, i_k}$ and $\pi_{jl} = \sum_{i_j=i_l=1} p_{i_1, \dots, i_k}$.

Assume that we have a prior on the parameter space. Here we consider the conjugate prior

$$\pi(p) = I(0 < \sum_{j=1}^k i_j \leq s) \frac{\Gamma(\sum_{i_j=0 \text{ or } 1} \alpha_{i_1, \dots, i_k})}{\prod_{i_j=0 \text{ or } 1} \Gamma(\alpha_{i_1, \dots, i_k})} \prod_{i_j=0 \text{ or } 1} p_{i_1, \dots, i_k}^{\alpha_{i_1, \dots, i_k}}, \quad (2)$$

which is a Dirichlet distribution with $C_s^k(2^s - 1)$ parameters.

Under this setup, we have the posterior distribution

$$\begin{aligned} \pi(p|n^*) &= f(n^*|p)\pi(p) \\ &= \frac{\Gamma(\sum_{i_j=0 \text{ or } 1} (\alpha_{i_1, \dots, i_k} + n_{i_1, \dots, i_k}))}{\prod_{i_j=0 \text{ or } 1} \Gamma(\alpha_{i_1, \dots, i_k} + n_{i_1, \dots, i_k})} \prod_{i_j=0 \text{ or } 1} p_{i_1, \dots, i_k}^{\alpha_{i_1, \dots, i_k} + n_{i_1, \dots, i_k}}. \end{aligned} \quad (3)$$

Through the form of the posterior distribution, we can derive the Bayes estimator for each p_{i_1, \dots, i_k} under the squared error loss function. The Bayes estimator $\hat{\pi}_j$ of π_j is equal to the summation of the Bayes estimator of p_{i_1, \dots, i_k} , where $i_j = 1$. We can base this on the Bayes estimators of π_j to rank the significance of π_j . However, if the ranking is based only on the Bayes estimators, it may lack of enough confidence to convince people to accept the ranking result. Therefore, establishing a multiple testing approach under a specific tolerance error to certify that the resulting rank is accurate is an essential issue.

2.1 Prior selection

To decide the Dirichlet prior (2), we have to select appropriate values for the parameters. If the empirical experience has provided us the model of the prior, then we can directly use this prior. If we do not have a specific prior but have the past data of the survey for this multiple responses question, we can choose an appropriate prior based on the data. When the past data is complete, meaning that it has the records for the number of each $C_s^k(2^s - 1)$ possible answer, we can set

the value of the parameter $\alpha_{i_1 \dots i_k}$ in the Dirichlet distribution to be $n(m_{i_1 \dots i_k})/m$, where m denotes the number of respondents and $m_{i_1 \dots i_k}$ denotes the number of respondents selecting the responses v_h and not selecting $v_{h'}$ if $i_h = 1$ and $i_{h'} = 0$ for the past data. In this way, the sum of $\alpha_{i_1 \dots i_k}$ is equal to n , which leads to the equal contribution of the past data and the current survey data. This equal weight contribution can balance the past information and the current survey data in the statistical inference.

On the other hand, the past data may be incomplete. It could only have the records of the number of each responses selected, but not the number of each $C_s^k(2^s - 1)$ possible answer selected. In this case, it is hard to estimate all parameters for the prior. Instead of estimating each parameter directly, for a response, we can set the equal weight to the parameters $\alpha_{i_1 \dots i_k}$ in the prior with corresponding $p_{i_1 \dots i_k}$ selecting this response. Then take the sum of weights assigned to the answer to be the value of parameter.

The selection of a prior may be a key issue for analyzing the data. The approach for prior estimation suggested above may not approximate the true prior very well. Thus, to achieve more accurate estimation, a more careful investigation to determine the prior selection is necessary. Since this study does not focus on prior selection, we do not provide a comprehensive discussion of this issue herein.

3 Testing approach

3.1 Multiple testing

In this section, we propose several multiple testing methods to make comparison of π_j . Assume that there are k responses and we are interested in testing

$$\begin{aligned} H_{01} : \pi_2 &\leq \pi_1 \text{ vs } H_{11} : \pi_2 > \pi_1 \\ H_{02} : \pi_3 &\leq \pi_2 \text{ vs } H_{12} : \pi_3 > \pi_2 \\ &\dots \\ H_{0k-1} : \pi_k &\leq \pi_{k-1} \text{ vs } H_{1k-1} : \pi_k > \pi_{k-1}. \end{aligned} \tag{4}$$

Note that it may be reasonable to test the equality of $\pi_1 = \dots = \pi_k$ first, and then proceed to test the one-sided test when the equality of $\pi_i, i = 1, \dots, k$ is rejected. The approach for testing a point null hypothesis is discussed by Berger (1985). It is necessary to assign a probability ξ_0 to the case $H_0 : \pi_1 = \dots = \pi_k$ and spreading out the probability of $1 - \xi_0$ on the alternative hypothesis H_0^c . Since the probability that the $\pi_i, i = 1, \dots, k$ are equal may be low, we do not investigate testing the point null hypothesis in depth in this study. In addition, according to the ranking criterion (9) used in this study, for ranking two responses π_i and π_j , both one-sided hypotheses $H_0 : \pi_i > \pi_j$ and $H_0 : \pi_j > \pi_i$ are considered, which may reflect the information obtained from the point null hypothesis.

For testing (4), the decision rules considered here is to control the posterior expected false discovery rate. The concept of false discovery rate (FDR) was proposed by Benjamini and Hochberg (1995) to determine optimal thresholds under this criterion in a multiple testing setting. When we test multiple hypotheses, the possible outcomes (over the l tests) may be summarized in Table 1.

Table 1. Outcomes of multiple tests. The notations l is the total number of hypotheses, l_0 is the unknown number of the true null hypotheses, l_1 is the unknown number of the false null hypotheses, V is the number of false positives, T is the number of false negatives, S is the number of rejected null hypotheses that are false, U is the number of rejected null hypotheses that are true and D is the number of rejected null hypotheses.

	Test result		
	number of H_{0i} accepted	number of H_{0i} rejected	
real state			
number of true H_{0i}	U	V	l_0
number of false H_{0i}	T	S	l_1
	$l - D$	D	l

We define the false discovery rate, posterior false discovery rate, false negative rate and posterior false negative rate for the frequentist and Bayesian setting based on the literature as follows.

First, some notations and definitions are given. Let z_i denote an indicator that the i th hypothesis H_{0i} is false and $v_i = P(z_i = 1 | n^*)$ denote the marginal posterior probability of $\pi_{i+1} > \pi_i$. The rejection of the H_{0i} is denoted by $d_i = 1$, otherwise $d_i = 0$. Let $z = (z_1, \dots, z_{k-1})$ and $d = (d_1, \dots, d_{k-1})$. Under the frequentist setup, the false discovery rate and false negative rate are denoted by the expectations $E[\frac{V}{D+\epsilon}]$ and $E[\frac{T}{n-D+\epsilon}]$ respectively, where $D = \sum d_i$ and ϵ is a small constant to avoid a zero denominator.

Let

$$FDR(d, z) = \frac{\sum d_i(1 - z_i)}{D + \epsilon}$$

and

$$FNR(d, z) = \frac{\sum (1 - d_i)z_i}{n - D + \epsilon}$$

Under a Bayesian setting, these error rates are defined as the posterior expected false discovery rate denoted by $\overline{FDR}(d, n^*)$ and the posterior expected false neg-

ative rate denoted as $\overline{FNR}(d, n^*)$, where

$$\overline{FDR}(d, n^*) = \int FDR(d, z) dp(z|n^*) = \frac{\sum d_i(1 - v_i)}{D + \epsilon}$$

and

$$\overline{FNR}(d, n^*) = \int FNR(d, z) dp(z|n^*) = \frac{\sum (1 - d_i)v_i}{n - D + \epsilon}.$$

The posterior expected false discovery count $\overline{FD}(d, n^*)$ and the posterior expected false negative count $\overline{FN}(d, n^*)$ are defined as

$$\overline{FD}(d, n^*) = \sum d_i(1 - v_i)$$

and

$$\overline{FN}(d, n^*) = \sum (1 - d_i)v_i.$$

By definition, v_i in the model for the multiple responses questionnaire can be expressed as

$$\propto \int \cdots \int (I(0 < \sum_{j=1}^k i_j \leq s) I(\pi_{l+1} > \pi_l) \frac{\Gamma(\sum_{i_j=0 \text{ or } 1} \alpha_{i_1 \dots i_k} + n_{i_1 \dots i_k})}{\Gamma(\alpha_{i_1 \dots i_k} + n_{i_1 \dots i_k})} \prod_{i_j=0 \text{ or } 1} p_{i_1 \dots i_k}^{\alpha_{i_1 \dots i_k} + n_{i_1 \dots i_k}} \prod_{i_j=0 \text{ or } 1} dp_{i_1 \dots i_k}, \quad (5)$$

which may be difficult to derive directly because it is a multiple integration. Instead of deriving its exact value, we can approximate it by simulation or using the normal approximation.

Theorem 1. By the normal approximation, the multiple integration (5) can be approximated by

$$\Phi\left(\frac{B}{\sqrt{C}}\right),$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution,

$$A = \sum_{i_j=0 \text{ or } 1} (\alpha_{i_1 i_2 \dots i_k} + n_{i_1 i_2 \dots i_k}),$$

$$B = \frac{\sum_{i_{l+1}=1, i_l=0} (\alpha_{i_1 i_2 \dots i_k} + n_{i_1 i_2 \dots i_k}) - \sum_{i_l=1, i_{l+1}=0} (\alpha_{i_1 i_2 \dots i_k} + n_{i_1 i_2 \dots i_k})}{A}$$

and

$$\begin{aligned}
C &= \frac{1}{A^2(A+1)} \left(\sum_{i_{l+1}=1, i_l=0} (\alpha_{i_1 i_2 \dots i_k} + n_{i_1 i_2 \dots i_k})(A - \alpha_{i_1 i_2 \dots i_k} - n_{i_1 i_2 \dots i_k}) \right. \\
&+ \sum_{i_l=1, i_{l+1}=0} (\alpha_{i_1 i_2 \dots i_k} + n_{i_1 i_2 \dots i_k})(A - \alpha_{i_1 i_2 \dots i_k} - n_{i_1 i_2 \dots i_k}) \\
&+ 2 \sum_{i'_{l+1}=1, i'_l=0} \sum_{i''_l=1, i''_{l+1}=0} (\alpha_{i'_1 i'_2 \dots i'_k} + n_{i'_1 i'_2 \dots i'_k})(\alpha_{i''_1 i''_2 \dots i''_k} + n_{i''_1 i''_2 \dots i''_k}) \Big). \quad (6)
\end{aligned}$$

The proof is given in the Appendix.

3.2 Testing procedures

We will introduce several multiple testing procedures in Berger (1985) and Muller et al (2004) for testing (4).

Method 1. The decision of accepting or rejecting the null hypothesis is based on the specific loss function proposed by Berger (1985), which is defined as

$$\begin{cases} 0 & \text{if the decision taken is right} \\ c & \text{if we reject } H_{0i} \text{ when it is true} \\ 1 & \text{if we accept } H_{0i} \text{ when it is false} \end{cases} \quad (7)$$

where $c(\geq 0)$ and 1 represent the losses for making a wrong decision due to a false positive and a false negative error, respectively. In this criterion, the loss function can be written as

$$L_N(d, n^*) = c\overline{FD} + \overline{FN}.$$

Method 2. The second method is to consider the loss function

$$L_R(d, n^*) = c\overline{FDR} + \overline{FNR}.$$

Method 3. We also consider bivariate loss functions that explicitly acknowledge the two competing goals, leading to the following posterior expected losses:

$$L_{2R}(d, n^*) = (\overline{FDR}, \overline{FNR}).$$

We can define the optimal decisions under L_{2R} as the minimization of \overline{FNR} subject to $\overline{FDR} \leq \alpha_{2R}$.

By Muller et al (2004), under the three loss functions, the optimal decision that minimizes the loss functions takes the form

$$d_i = I(v_i \geq t^*), \quad (8)$$

where t^* are $t_N^* = c/(c+1)$, $t_R^*(n^*) = v_{(n-D^*)}$ and $t_{2R}^*(n^*) = \min\{s : FDR(s, n^*) \leq \alpha_{2R}\}$ under the loss functions L_N, L_R and L_{2R} , respectively. In the expressions for t_R^* and t_{2R}^* , $v_{(i)}$ is the i th order statistic of $\{v_1, \dots, v_n\}$, and D^* is the optimal number of discoveries that is found by minimizing the function (A.1) in Muller et al. (2004).

A simulation study for comparing the three methods is given in Section 5.

4 Ranking approach and ranking consistency

If not all of the hypotheses in (4) are rejected, there does not have enough evidence to rank all responses. An objective way to rank the responses is to test the hypothesis $\pi_i > \pi_j$ for each i and j . There are totally C_2^k hypotheses for k responses. The rank of the i th response can be defined as follows.

$$R_i = k - \sum_{j=1, j \neq i}^k I(\pi_i > \pi_j). \quad (9)$$

Using the criterion (9), we define a response the most significant if it has smallest R_i value and we rank it first. The response with second smallest R_i value is defined to be the second significant response and so on.

By Wang (2008a), a reasonable ranking approach may need to satisfy the ranking consistency property. The property is modified here to fit the Bayesian set up as follows.

Bayesian Ranking consistency property:

A test is called ranking consistent if $\pi_j = \pi_i$ is rejected by the test, then $\pi_j = \pi_g$

should also be rejected by the test with the same level if the Bayes estimator of $I_{\pi_j - \pi_i > 0}$ is less than the Bayes estimator of $I_{\pi_j - \pi_g > 0}$.

From the examples given in Wang (2008a), under the frequentist framework, the tests derived by the conventional approaches do not possess the property of frequentist ranking consistency. It is still unknown if there exist ranking consistent tests under the frequentist framework. When considering the problem under the Bayesian framework, it is easier to find the ranking consistent tests.

Theorem 2. The three testing procedures (8) considered in Section 3 for different t^* values under the loss functions L_N , L_R and L_{2R} , respectively are ranking consistent.

Proof. For the three tests in Section 3, the decision rules of the tests are based on (8). From the form, for a fixed cutoff t^* , the decision rule only depends on the Bayes estimator v_i of H_{0i} . If a hypothesis H_{0i} with a smaller v_i is rejected, then a hypothesis H_{0j} with a larger v_j is accordingly rejected by the rule. Thus, the proof is complete.

4.1 Penalty Score

In this section, we will set up a penalty score to evaluate the three methods from the viewpoint of ranking error. To rank the i th responses, for a method, we need to calculate their R_i values using this method, then use the value R_i to rank the responses. A penalty score is defined as the summation of the absolute values of the true rank and the rank derived by the method. For example, in the case of $k = 5$, if the true rank of the first response is 1, and the true rank of the second response is 2, etc. We use the notation $(1, 2, 3, 4, 5)$ to denote the true rank. If the rank derived by a method for an observation is $(2, 1, 3, 5, 4)$, the penalty score for the method given by the observation is $|1 - 2| + |2 - 1| + |3 - 3| + |4 - 5| + |5 - 4| = 4$. We conduct a simulation for 1000 replicates to compare the expected penalty scores for the three methods. The simulation procedure is as follows.

Step 1. First we set up a prior for α .

Step 2. We generate a set of p from the prior distribution with respect to the α value in Step 1. From this p , we can attain a true rank for π based on this p .

Step 3. Using the pmf function (1) with the p in Step 2, we generate a set n^* .

Step 4. We set up the $\binom{k}{2}$ null hypotheses for any two different π_i . Then based on n^* in Step 3, we calculate the Bayes estimator of the indicator function of each hypothesis. Then apply the three methods in Section 3 to test each null hypothesis. Use (9) to rank the k responses and calculate the penalty score from the derived rank and the true rank in Step 2 for each method.

Step 5. Repeat Steps 2-4 1000 times. Take the average of the penalty score in Step 4 for each method and the approximated expected penalty score for each method is derived.

Remark 1. When we consider the testing for the $2C_2^k$ hypotheses to calculate the $R_i, i = 1, \dots, k$, although the hypotheses are not exactly in the form of (4), we can still apply the methods for the general testing of the $2C_2^k$ hypotheses.

5 Simulation result

5.1 Rejection rate

A simulation study is conducted to evaluate the performance of the three methods in this section. We first set up a known prior of the form (2) on the parameter space. Let $w_j = \sum_{i_j=1} \alpha_{i_1 \dots i_k}, j = 1, \dots, k$. A condition for the prior setting is given in Section 2.1. The simulation procedure is to generate a set of p . Then use the p in (1) to generate a set of n^* . Next, calculate v_i conditioning on the n^* and use the v_i for the three different loss criteria. For testing the $k - 1$ hypotheses of (4), we can count the rejection number for the $k - 1$ hypotheses for the three methods. Although the truth of the $k - 1$ hypotheses depends on p , by the property of the Dirichlet distribution, we have $E(\pi_i) < E(\pi_j)$ if $w_i < w_j$. If we repeat the simulation procedure many times, the rejection number for the hypothesis

$H_{0i} : \pi_{i+1} < \pi_i$ of a good testing should be close to $P(\pi_{i+1} > \pi_i)$. Thus, we can use the criterion to evaluate the testing methods. We repeat the simulation process 1000 times and the results are shown in Examples 1 and 2. The t^* values for the first two tests in these examples are selected such that their corresponding c values in L_N and L_R can minimize penalty score presented in Section 5.2. The t^* values for the third test in these examples are selected such that their corresponding α_{2R} values in L_{2R} can minimize the penalty score presented in Section 5.2.

Example 1. Consider the case of $k = 5$ and a Dirichlet prior distribution on the parameter space with $\alpha_{00000} = 0, \alpha_{00001} = 98, \alpha_{00010} = 63, \alpha_{00100} = 42, \alpha_{01000} = 28, \alpha_{10000} = 28$ and the others are equal 7. In this case, $w_1 = 133 = w_2 = 133 < w_3 = 147 < w_4 = 168 < w_5 = 203$. Under this setup, we have $P(\pi_2 > \pi_1) = 0.500, P(\pi_3 > \pi_2) = 0.859, P(\pi_4 > \pi_3) = 0.930$ and $P(\pi_5 > \pi_4) = 0.986$. To testing (4), we compare the three methods introduced in Section 3. The rejection rates for each method are listed in Table 2, where c values are 1 and 0.33 for L_N and L_R and α_{2R} value is 0.15 for L_{2R} .

Table 2: The rejection rates of the three methods corresponding to each hypothesis in (4) for 1000 replicates.

	$H_{01} : \pi_2 \leq \pi_1$	$H_{02} : \pi_3 \leq \pi_2$	$H_{03} : \pi_4 \leq \pi_3$	$H_{04} : \pi_5 \leq \pi_4$
L_N	0.473	0.939	0.981	1
L_R	0.188	0.798	0.951	0.996
L_{2R}	0.415	0.899	0.965	0.998

Example 2. Consider the case of $k = 5$ and a Dirichlet prior distribution on the parameter space with $\alpha_{00000} = 0, \alpha_{00001} = 56, \alpha_{00010} = 49, \alpha_{00100} = 42, \alpha_{01000} = 35, \alpha_{10000} = 70$ and the others are equal 7. In this case, $w_2 = 140 < w_3 = 147 < w_4 = 154 < w_5 = 161 < w_1 = 170$. We have $P(\pi_2 > \pi_1) = 0.006, P(\pi_3 > \pi_2) = 0.703, P(\pi_4 > \pi_3) = 0.696$ and $P(\pi_5 > \pi_4) = 0.688$. To test (4), we compare the three methods introduced in Section 3. The rejection rates for each method are

listed in Table 3, , where c values are 1 and 0.54 for L_N and L_R and α_{2R} value is 0.2 for L_{2R} .

Table 3: The rejection rates of the three methods corresponding to each hypothesis in (4) for 1000 replicates.

	$H_{01} : \pi_2 \leq \pi_1$	$H_{02} : \pi_3 \leq \pi_2$	$H_{03} : \pi_4 \leq \pi_3$	$H_{04} : \pi_5 \leq \pi_4$
L_N	0	0.754	0.759	0.747
L_R	0.001	0.903	0.911	0.907
L_{2R}	0	0.582	0.570	0.565

From Tables 2 and 3, the performance of the method under the loss function L_R seems worse than the other two methods because its rejection rate is not close to the probability of the indicator function of the alternative hypothesis in most cases. In Example 1, Method 3 seems better than Method 1. However, in Example 2, Method 1 is better than Method 3. Different situations result in different performances by these two methods. Overall, Method 1 and Method 3 may be superior to the Method 2 in many cases as shown in the simulation study.

Following the above procedures, the approximate expected score for the three methods can be derived. Note that the scores for Methods 1 and 2 depend on the value of c , and the score for Method 3 depends on the value of α_{2R} . In the real application, the selection of c in Methods 1 and 2 may depend on the true cost of the wrong decision making and the selection of α_{2R} in Method 3 may depend on the allowed tolerance error. However, from a theoretical viewpoint, we can investigate the situation of c and α_{2R} such that the three methods have the smallest penalty score.

Based on the simulation procedures, the performances of the expected penalty score for different c and α_{2R} corresponding to $\alpha_{i_1 \dots i_k}$ in Examples 1 and 2 are presented in Figures 1 and 2.

The minimum expected penalty scores for Methods 1-3 are 1.162, 1.646 and

1.173 in Figure 1, which occurs at $c = 1, c = 0.33$ and $\alpha_{2R} = 0.15$, respectively. The minimum expected penalty scores for Methods 1-3 are 2.274, 3.012 and 2.534 in Figure 1, which occurs at $c = 1, c = 0.54$ and $\alpha_{2R} = 0.2$, respectively. Basically, Figures 1 and 2 show that Method 1 has the smallest minimum expected penalty scores, followed by Method 3. Method 2 has the largest minimum expected penalty scores, which leads to the worst performance among these three methods. From the viewpoint of ranking, this consequence coincides with the results in Section 5.1.

6 A Real Data Example

In this section, we use a real data example to illustrate the methods and present a case which it is ranking inconsistent under the frequentist framework, but is ranking consistent under the Bayesian framework. This example is a survey of 49609 first-year college students in Taiwan about their preferences in their college study. The data set can be accessed at <http://srda.sinica.edu.tw> and it is available upon request from the first author. We list one of the multiple responses questions in the questionnaire as follows.

Question: What kind of experience do you expect to receive during the period of college study? (Select at least one response)

1. Read over the Chinese and foreign classic literature
2. Travel around Taiwan
3. Present academic papers in conferences
4. Lead large-scale activities
5. Be on a school team
6. Be a cadre of student associations
7. Participate internship programs
8. Fall in love
9. Have sexual experience
10. Travel around the world

11. Make many friends

12. Others

We are interested in ranking the responses of this multiple responses question according to students' preference. To make a clear illustration, we do not consider the problem of ranking all responses, but the problem of ranking the five responses: read Chinese and foreign classics, present academic papers in conferences, lead large-scale activities, be on a school team and be a student association cadre.

The population of the survey is the whole data set including 49609 interview data. Since we have all data, we can obtain the true ranks of the five responses. To illustrate the methods, suppose that we do not have the whole data set, but only have the interview data of 100 randomly selected respondents. Note that from the whole data set, the numbers of respondents selecting the five responses are 8858, 5358, 10578, 6823 and 12145. The first, second and third ranks show that the students prefer to "be a student association cadre", "lead large-scale activities" and "read Chinese and foreign classics".

In this example, the notations $i_1 = 1, i_2 = 1, i_3 = 1, i_4 = 1$ and $i_5 = 1$ in $n_{i_1 i_2 i_3 i_4 i_5}$ correspond to selection of the response "read Chinese and foreign classics", "present academic papers in conferences", "lead large-scale activities", "be on a school team" and "be a cadre of student associations", respectively.

According to a 100 randomly selected data, we have $n_{10000} = 19, n_{01000} = 5, n_{00100} = 7, n_{00010} = 6, n_{00001} = 10, n_{11000} = 3, n_{10100} = 0, n_{10010} = 0, n_{10001} = 5, n_{01100} = 1, n_{01010} = 0, n_{01001} = 1, n_{00110} = 0, n_{00101} = 8, n_{00011} = 2, n_{11100} = 0, n_{11010} = 1, n_{11001} = 0, n_{01110} = 0, n_{01101} = 3, n_{01011} = 0, n_{00111} = 8, n_{10110} = 0, n_{10101} = 7, n_{10011} = 0, n_{11110} = 0, n_{11101} = 3, n_{11011} = 1, n_{10111} = 3, n_{01111} = 3$ and $n_{11111} = 4$ for the 100 data. This leads to $m_1 = 46, m_2 = 24, m_3 = 47, m_4 = 29, m_5 = 58$. From the data, the most selected responses is "be a cadre of student associations". Next is "lead large-scale activities", followed by "read Chinese and foreign classics". Consequently, we have $m_{(5)} = 58, m_{(4)} = 47, m_{(3)} = 46, m_{(2)} =$

29, $m_{(1)} = 24$. Now we are interested in testing

$$\begin{aligned} H_{01} : \pi_{(5)} &\leq \pi_{(4)} \text{ vs } H_{11} : \pi_{(5)} > \pi_{(4)} \\ H_{02} : \pi_{(5)} &\leq \pi_{(3)} \text{ vs } H_{12} : \pi_{(5)} > \pi_{(3)}. \end{aligned} \tag{10}$$

In this case, the likelihood ratio test does not lead to the rejection of the hypotheses. Thus, we use the Wald and generalized score tests to illustrate the ranking inconsistency property. When testing H_{01} , the values of the two test statistics with respect to the Wald test and generalized score test under the frequentist framework are 2.17 and 2.12. The upper 0.025 cutoff point of the standard normal distribution is 1.96, resulting in the rejection of H_{01} by the two tests with 0.025 type I error. However, when testing H_{02} , the values of statistics corresponding to the Wald test and generalized score test are 1.59, and 1.57, which does not lead to the rejection of H_{02} in both two tests. Since $|\pi_{(5)} - \pi_{(3)}| > |\pi_{(5)} - \pi_{(4)}|$, the above result leads to ranking inconsistency for the Wald and score tests under the frequentist framework.

Now we consider the Bayesian framework and implement Method 1, Method 2 and Method 3 for this example. According to the whole data, we assume a prior for $p_{i_1 i_2 i_3 i_4 i_5}$, which corresponds to $\alpha_{10000} = 13, \alpha_{01000} = 4, \alpha_{00100} = 8, \alpha_{00010} = 5, \alpha_{00001} = 11, \alpha_{11000} = 3, \alpha_{10100} = 2, \alpha_{10010} = 1, \alpha_{10001} = 3, \alpha_{01100} = 1, \alpha_{01010} = 0, \alpha_{01001} = 1, \alpha_{00110} = 1, \alpha_{00101} = 10, \alpha_{00011} = 3, \alpha_{11100} = 0, \alpha_{11010} = 0, \alpha_{11001} = 0, \alpha_{01110} = 0, \alpha_{01101} = 2, \alpha_{01011} = 0, \alpha_{00111} = 6, \alpha_{10110} = 0, \alpha_{10101} = 3, \alpha_{10011} = 1, \alpha_{11110} = 0, \alpha_{11101} = 1, \alpha_{11011} = 0, \alpha_{10111} = 2, \alpha_{01111} = 1$ and $\alpha_{11111} = 4$.

In the real applications, we can estimate the prior or derive a prior from past experience.

For implementing Method 1 and Method 2, we select $c = 1$, and $\alpha_{2R} = 0.15$ corresponding to Method 1 and Method 2, resulting in $t^* = 0.5$ and 0.9917 with respect to the two methods. For testing (10) under the given prior, we have $v_1 = 0.9919$ and $v_2 = 0.9947$. Consequently, by (8), H_{01} and H_{02} are both rejected

by the two methods.

In this case, the results show that the data leads to the conventional tests under the frequentist framework is ranking inconsistent, and the proposed methods are ranking consistent under the Bayesian framework.

7 Conclusion

Several methods for ranking the responses in a multiple responses question under Bayesian framework are proposed in this paper. The specified ranking criterion and ranking error penalty are established. Compared with the methods under the frequentist framework, these methods are more convincing because they have the property of Bayesian ranking consistency.

From the simulation study, the the methods using the loss functions L_N and L_{2R} are better than the method using the loss function L_R if we consider the cases where c and α_{2R} are selected such that the minimum expected penalty score occurs. However, in real applications, the selection of the constant c in L_N and L_R may depend on the economic cost. The same is true in the selection of α_{2R} . Since α_{2R} provides a tolerance error of false discovery rate, in real applications, the setup of α_{2R} may depend on the allowed tolerance error. For that reason, researchers may be dependent on cost when selecting the most useful approaches.

8 Appendix

Proof of Theorem 1. Note that $\pi_{l+1} - \pi_l = (\pi_{l+1} - \pi_{(l+1)l}) - (\pi_l - \pi_{(l+1)l})$. For a given n^* , let $A = \sum_{i_j=0 \text{ or } 1} (\alpha_{i_1 i_2 \dots i_k} + n_{i_1 i_2 \dots i_k})$. From the property of the Dirichlet distribution, we have the expectation and variance of $p_{i'_1 i'_2 \dots i'_k}$ equal to

$$\frac{(\alpha_{i'_1 i'_2 \dots i'_k} + n_{i'_1 i'_2 \dots i'_k})}{A}$$

and

$$\frac{(\alpha_{i'_1 i'_2 \dots i'_k} + n_{i'_1 i'_2 \dots i'_k})(A - \alpha_{i'_1 i'_2 \dots i'_k} - n_{i'_1 i'_2 \dots i'_k})}{A^2(A + 1)}$$

respectively.

The covariance of $p_{i'_1 i'_2 \dots i'_k}$ and $p_{i''_1 i''_2 \dots i''_k}$ is equal to

$$\frac{-(\alpha_{i'_1 i'_2 \dots i'_k} + n_{i'_1 i'_2 \dots i'_k})(\alpha_{i''_1 i''_2 \dots i''_k} + n_{i''_1 i''_2 \dots i''_k})}{A^2(A+1)}.$$

Therefore, from the above facts and straightforward calculation, the expectation of $\pi_{l+1} - \pi_l$ can be rewritten as

$$\begin{aligned} E(\pi_{l+1} - \pi_l) &= E((\pi_{l+1} - \pi_{(l+1)l}) - (\pi_l - \pi_{(l+1)l})) \\ &= B \end{aligned}$$

and the variance of $\pi_{l+1} - \pi_l$ can be rewritten as

$$\begin{aligned} Var(\pi_{l+1} - \pi_l) &= Var((\pi_{l+1} - \pi_{(l+1)l}) - (\pi_l - \pi_{(l+1)l})) \\ &= Var(\pi_{l+1} - \pi_{(l+1)l}) + Var(\pi_l - \pi_{(l+1)l}) - 2Cov((\pi_{l+1} - \pi_{(l+1)l})(\pi_l - \pi_{(l+1)l})) \\ &= C \end{aligned} \tag{11}$$

By normal approximation, we have

$$\begin{aligned} v_l &= P(\pi_{l+1} - \pi_l > 0) \\ &= P\left(\frac{\pi_{l+1} - \pi_l - E(\pi_{l+1} - \pi_l)}{\sqrt{Var(\pi_{l+1} - \pi_l)}} > \frac{-E(\pi_{l+1} - \pi_l)}{\sqrt{Var(\pi_{l+1} - \pi_l)}}\right) \\ &= P\left(Z > \frac{-E(\pi_{l+1} - \pi_l)}{\sqrt{Var(\pi_{l+1} - \pi_l)}}\right) \\ &= \Phi\left(\frac{E(\pi_{l+1} - \pi_l)}{\sqrt{Var(\pi_{l+1} - \pi_l)}}\right) \end{aligned} \tag{12}$$

The proof is complete.

Acknowledgements: The authors thank the editor and references for helpful comments.

References

- [1] Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rates: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* (57), 289V300.
- [2] Agresti, A. and Liu, I.M. (1999) Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics* 55, 936-943.
- [3] Agresti, A. Liu, I.M. (2001) Strategies for modeling a categorical variable allowing multiple category choices. *Sociological Methods and Research* 29, 403V434.
- [4] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.
- [5] Bilder, C. R., Loughin, T. M. and Nettleton, D. (2000) Multiple marginal independence testing for pick any/c variables. *Comm.Statist.Simulation Comput.*, 29(4), 1285-1316.
- [6] Decady, Y. J. and Thomas, D. H. (2000). A simple test of association for contingency tables with multiple column responses. *Biometrics* 56, 893-896.
- [7] Gopalan, R. and Berry, D. A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association* 93, 1130V1139.
- [8] Do, K., Muller, P. and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *J. R. Stat. Soc. C.*, 54, 627V644.

- [9] Pammer, S., Fong, D. K. H. and Arnold, S. F. (2000). Forecasting the Penetration of a New Product: A Bayesian Approach. *Journal of Business and Economic Statistics*, 18, no. 4, 428-435.
- [10] Gonen, M., Westfall, P. H. and Johnson, W. O. (2003). Bayesian Multiple Testing for Two-Sample Multivariate Endpoints. *Biometrics*, 59, 76-82.
- [11] Loughin, T. M. and Scherer, P. N. (1998). Testing for association in contingency tables with multiple column responses. *Biometrics* 54, 630-637.
- [12] Muller P, Parmigiani G, and Rice K. (2007). "FDR and Bayesian decision rules." In *Bayesian Statistics 8.* (Bernardo, J. et al. ed.) Oxford University Press.
- [13] Miranda-Moreno, L. F., Labbe, A. and Fu, L. (2007). Bayesian multiple testing procedures for hotspot identification. *Accident Analysis and Prevention*, 39, 1192V1201.
- [14] Muller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association*, 99, no.468, 990-1001.
- [15] Scott, J. (2009). "Nonparametric Bayesian multiple testing for longitudinal performance stratification." *Annals of Applied Statistics*.
- [16] Scott, J.G. and Berger, J.O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Stat. Plann. Inference* 136, no. 7, 2144V2162.
- [17] Umesh, U. N. (1995). Predicting nominal variable relationships with multiple responses. *Journal of Forecasting* 14, 585-596.
- [18] Wang, H. (2008a). Ranking responses in multiple responses questions. *Journal of Applied Statistics*, 35, 465-474.

- [19] Wang, H. (2008b) Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *Journal of Multivariate Analysis*, 99, 896-911.



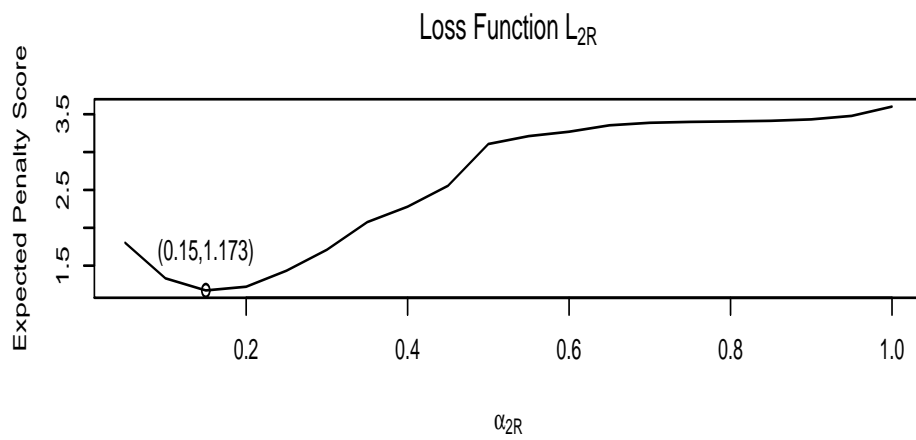
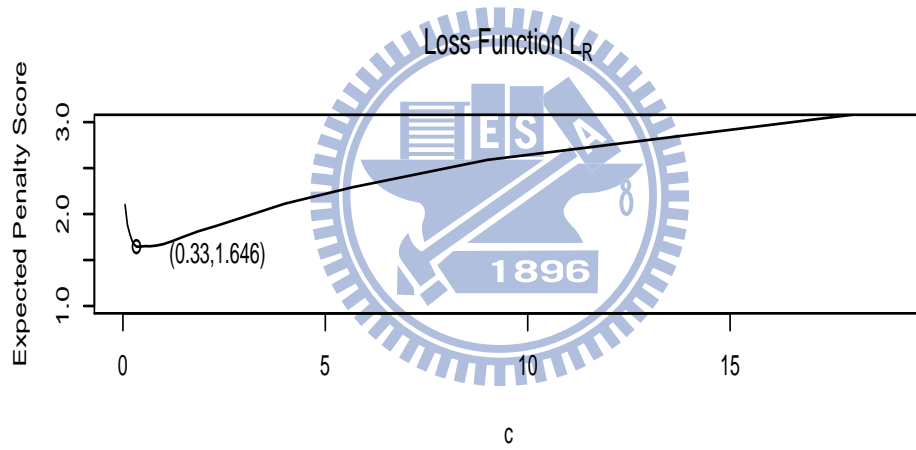
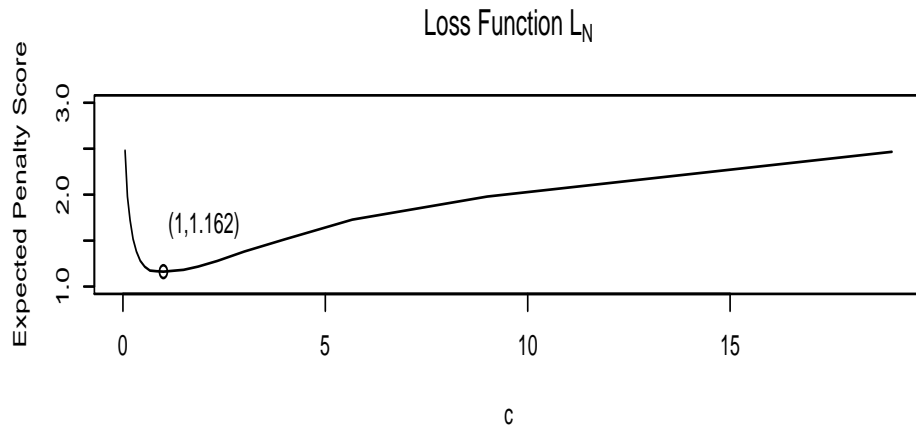


Figure 1: The expected penalty scores of the three methods under the condition of Example 1

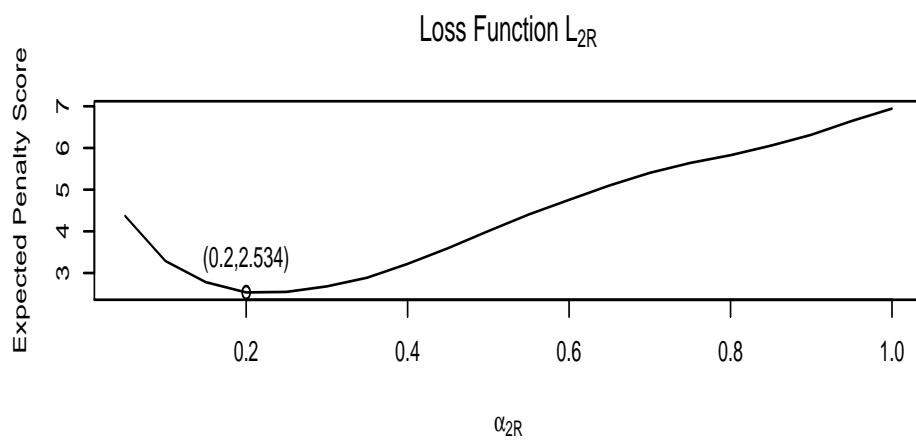
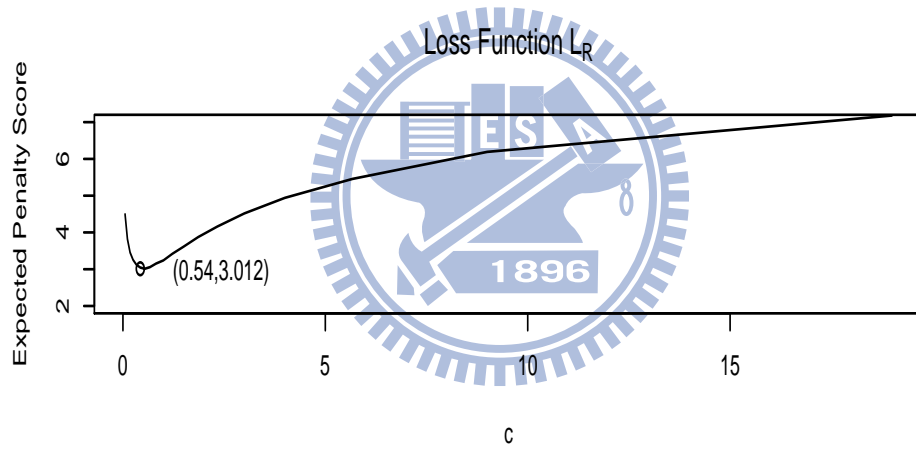
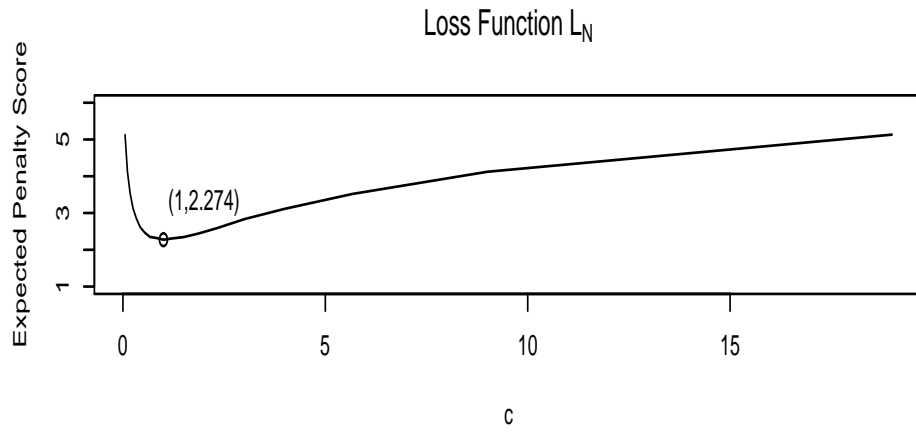


Figure 2: The expected penalty scores of the three methods under the condition of Example 2