Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

*J. R. Statist. Soc.* A (2014)
**177**, *Part* 1, *pp.* 191–208

# Bayesian ranking responses in multiple-response questions

Hsiuying Wang and Wei Heng Huang

*National Chiao Tung University, Hsinchu, Taiwan*

[Received August 2011. Final revision November 2012]

**Summary.** Questionnaires are important surveying tools that are used in numerous studies. Analyses of multiple-response questions are not as well established in detail compared with single-response questions. Wang has proposed several methods for ranking responses in multiple-response questions under the frequentist set-up. However, prior information may exist for ranks of responses in numerous situations. Therefore, establishing a methodology that combines updated survey data and past information for ranking responses is an essential issue in questionnaire data analysis. This study develops Bayesian ranking methods based on several Bayesian multiple-testing procedures to rank responses by controlling the posterior expected false discovery rate. Moreover, a simulation is conducted to compare these approaches, and a real data example is presented to show the effectiveness of the methods proposed.

*Keywords*: Bayes estimator; Dirichlet prior; Multiple-response question; Single-response questions; Surveys

## 1. Introduction

Questionnaires are a commonly used tool in numerous felds for collecting information, and they are especially important for marketing or management studies. There are two types of question: single-response questions and multiple-response questions. Response models that are related to this issue have been discussed in Thissen and Steinberg (1984, 1986).

The analyses of multiple-response questions are not as deeply established as those for single-response questions, and approaches for analysing multiple-response questions were inadequate until recently (Umesh, 1995; Loughin and Scherer, 1998; Decady and Thomas, 1999; Bilder *et al.*, 2000; Agresti and Liu, 1999, 2001).

These studies have mainly focused on analysing the dependence between a single-response question and a multiple-response question. In practice, the majority of researchers are interested in ranking the responses to questions according to the probability of the response being chosen. This ranking response issue may be the primary interest of survey analysis; thus, the issue of ranking responses in a multiple-response question is the focus of this study.

We discuss the problem by the example provided by Wang (2008a); a company designs a marketing survey to assist in the development of an insect killer. Several factors that can influence sales are examined by implementing a questionnaire, including high quality, price, packaging and smell. The researchers want to know the significance rank for these factors so that they can design a product with lower cost and higher profit.

A group of individuals are surveyed regarding purchasing an insect killer by completing

*Address for correspondence*: Hsiuying Wang, Institute of Statistics, National Chiao Tung University, Hsinchu 30010, Taiwan.
E-mail: wang@stat.nctu.edu.tw

 0964–1998/14/177191

questionnaires. A multiple-response question from the questionnaire is as follows.

> 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) other.

Wang (2008a) proposed several approaches for solving this problem under the frequentist set-up, i.e. the responses are ranked on the basis of only the current survey data. However, empirical information may exist in real applications for the probabilities of the responses being chosen. In this case, the Bayesian approach is more appropriate for analysing the data, which can associate the current survey data with past information. The latter case is the focus of this study, and this ranking issue is considered under the Bayesian framework.

For example, the British Household Panel Survey is a longitudinal study of individuals who were living in private households in Great Britain. For a longitudinal survey, the current survey data may be associated with past surveys, and the prior information may be available from previous interviews. Examples and discussions regarding the panel survey are referred to in Jenkins *et al.* (2006) and Jäckle and Lynn (2007) and Lynn *et al*. (2012). Related Bayesian applications are referred to in Pammer *et al.* (2000).

Under the Bayesian framework, we assume that prior information regarding the parameter space is available, and we rank the responses on the basis of a survey study and the prior information. For a single-response question, respondents who select different responses are independent; as a result the response ranking issue is associated with the usual Bayesian multiple-testing problem that was presented by Muller *et al.* (2004). However, respondents who select different responses in a multiple-response question may be dependent. Handling dependent data is more challenging than handling independent data; thus, analysing multiple-response questions is complex. In this paper, approaches based on Bayesian multiple-testing procedures are proposed for ranking the responses to multiple-response questions.

These methodologies are an extension of the methods in Muller *et al.* (2004), who proposed several criteria for Bayesian multiple testing. Miranda-Moreno *et al.* (2007) applied the method from the study of Muller *et al.* (2004) to identify hot spots in engineering. Wang (2008b) involved estimating the proportions in a multinomial distribution. Further details regarding Bayesian multiple testing and applications have been discussed by Gopalan and Berry (1998), Do *et al.* (2005), Gonen *et al.* (2003), Scott and Berger (2006), Muller *et al.* (2006, 2007) and Scott (2009). Moreover, related studies regarding Bayesian ranking methods are referred to in Berger and Deely (1988), where items are ranked on the basis of the posterior probability of the null hypothesis or the Bayes factor. Furthermore, Lin *et al.* (2006) proposed the loss function approach for ranking data. Although these methodologies provide rules for ranking, they do not create rules to evaluate ranking errors. This study proposes a method for evaluating the ranking error and applying the loss function method for the analysis of multiple-response questions.

The conventional Bayesian multiple-testing method involves calculating the posterior probability or the Bayes factor of the null hypothesis and rejecting or accepting the null hypothesis on the basis of these calculations. The criterion for rejecting the null hypothesis involves determining whether the posterior probability or the Bayes factor is larger compared with a specific critical value. Critical value selection in the conventional method is normally independent of observations and the sample size. When the sample size is large, the posterior probability can be used to reject or accept the null hypothesis with more confidence. When the sample size is more limited, a stricter critical value for the posterior probability may be required to avoid large false discovery rates. Because the conventional method does not assist in selecting a critical value based on observations, it cannot guarantee the identification of an appropriate decision. The false discovery rate procedure that has been proposed in the literature regarding Bayesian multiple

testing is adopted in this study. The false discovery rate procedure is a statistical method that is used for multiple-hypothesis testing and was designed to control the false discovery rate or false negative rate. Further disscussions regarding the Bayesian false discovery rate procedure are referred to in Muller *et al.* (2004). The Bayesian ranking property is also discussed in this study.

Wang (2008a) provided examples showing that the conventional testing approaches under the frequentist framework do not provide ranking-consistent results. The ranking consistency rule under the frequentist framework requires that the rank order of responses is consistent with the selection order for each response. This property is a reasonable criterion for confirming the validity of the testing approach. However, under the frequentist framework, satisfactory approaches for ranking responses with the property of ranking consistency have not yet been discovered. In this study, we propose a Bayesian ranking consistency rule because the ranking consistency rule under the frequentist approach cannot directly apply to the Bayesian framework. The methods proposed are shown to have the Bayesian ranking-consistent property.

This paper is organized as follows. A Bayesian model for multiple-response questions is discussed in Section 2. Several Bayesian multiple-testing procedures for testing the order of responses are proposed in Section 3. In Section 4, a ranking criterion for ranking responses is proposed, and the Bayesian multiple-testing procedures are shown to be consistent. Simulation studies for comparing the rejection rates of various methodologies are presented in Section 5. In addition, a real data example is provided in Section 6 for discussing the ranking consistency property. Finally, a conclusion is given in Section 7.

## 2. Model

First, question 1 from Section 1 is used to illustrate the model and is extended to the general model. In question 1 (multiple-response question), there are a total of $2^5 - 1 = 31$ possible answers because the case that respondents do not select any response is excluded. The 31 random variables constitute a multinomial distribution with multinomial proportions $p \in P = \{p_{i_1 i_2 i_3 i_4 i_5}, i_j = 0 \text{ or } i_j = 1 \text{ and } 0 < \Sigma_{j=1}^{5} i_j \leqslant 5\}$, where $i_j$ cannot be simultaneously equal to 0. The selection of at least one response is required for this multiple-response question. This requirement can prevent confusion with the missing value case. A questionnaire designer normally provides an 'other' response when designing a multiple-response question, which includes all other possible responses. This can prevent the scenario in which a respondent does not find any suitable responses. If a multiple-response question does not include all possible responses, there are two possible reasons for explaining why a respondent did not answer a question. One reason is because no response is suitable for the respondent. The other reason is because the respondent desires not to answer the question, which is the missing value case. There are various methods for handling these two situations. To avoid confusing these two situations, a better method involves including an other response and requiring respondents to select at least one response.

For the general case, assume that a multiple-response question has $k$ responses, $v_1, \ldots, v_k$, and we interview $n$ respondents. Each respondent is asked to choose at least one and at most $s$ answers for this question, where $0 < s \leqslant k$. If $s = 1$, it is a single-response question. There are a total of $c = C_1^k + \ldots + C_s^k$ possible kinds of answer from which respondents can choose, where $C_i^k = k!/\{i!(k - i)!\}$ is the number of ways of picking $i$ unordered outcomes from $k$ possibilities. Let $n_{i_1 \ldots i_k}$ denote the number of respondents selecting the responses $v_h$ and not selecting $v_{h'}$ if $i_h = 1$ and $i_{h'} = 0$. Note that $i_h = 1$ or $i_h = 0$ denotes that the $h$th response is selected or not selected respectively. And $p_{i_1 \ldots i_k}$ denotes the corresponding probability that a respondent selects the item $v_h$ and does not select the item $v_{h'}$ if $i_h = 1$ and $i_{h'} = 0$. For example, when $k = 7$, $n_{0100100}$ denotes the number of respondents selecting the second and the fifth responses and not selecting the other responses.

Thus, the probability function of $\mathbf{n}=\{n_{i_1\ldots i_k}, i_j=0 \text{ or } i_j=1 \text{ and } 0<\Sigma_{j=1}^{k} i_j \leqslant s\}$ is

$$f_s(\mathbf{n}|p)=I\left(0<\sum_{j=1}^{k} i_j \leqslant s\right) \frac{n!}{\prod\limits_{i_j=0 \text{ or } 1} n_{i_1\ldots i_k}!} \prod_{i_j=0 \text{ or } 1} p_{i_1\ldots i_k}^{n_{i_1\ldots i_k}}, \tag{1}$$

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

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

$$\eta(p)=I\left(0<\sum_{j=1}^{k} i_j \leqslant s\right) \frac{\Gamma\left(\sum\limits_{i_j=0 \text{ or } 1} \alpha_{i_1\ldots i_k}\right)}{\prod\limits_{i_j=0 \text{ or } 1} \Gamma(\alpha_{i_1\ldots i_k})} \prod_{i_j=0 \text{ or } 1} p_{i_1\ldots i_k}^{\alpha_{i_1\ldots i_k}}, \tag{2}$$

which is a Dirichlet distribution. In this study, the prior information is assumed to be known.

Under this set-up, we have the posterior distribution

$$\begin{aligned}\eta(p|\mathbf{n}) &= f_s(\mathbf{n}|p)\,\eta(p) \\ &= I\left(0<\sum_{j=1}^{k} i_j \leqslant s\right) \frac{\Gamma\left\{\sum\limits_{i_j=0 \text{ or } 1} (\alpha_{i_1\ldots i_k}+n_{i_1\ldots i_k})\right\}}{\prod\limits_{i_j=0 \text{ or } 1} \Gamma(\alpha_{i_1\ldots i_k}+n_{i_1\ldots i_k})} \prod_{i_j=0 \text{ or } 1} p_{i_1\ldots i_k}^{\alpha_{i_1\ldots i_k}+n_{i_1\ldots i_k}}.\end{aligned} \tag{3}$$

Through the form of the posterior distribution, we can derive the Bayes estimator for each $p_{i_1\ldots i_k}$ under the squared error loss function. The Bayes estimator $\hat{\pi}_j$ of $\pi_j$ is equal to the summation of the Bayes estimator of $p_{i_1\ldots i_k}$, where $i_j=1$. We can use the Bayes estimator of $\pi_j$ to rank the significance of $\pi_j$. Moreover, if we can associate a testing approach with the Bayes estimator to rank $\pi_j$, this method can lead to a more accurate result. Therefore, we intend to establish a multiple-testing approach under a specific tolerance error to improve the accuracy of the ranking result.

In real applications, we can obtain the prior information from past surveys. For example, we revisit the example of question 1. The researchers can obtain the prior information for this example from the past survey results about the marketing of an insect killer. To obtain a Dirichlet prior (2), if we have past data from the survey for this multiple-responses question, we can choose appropriate values for the parameters of prior (2) based on the data. Assume that $m$ respondents participated in the past survey and $n$ respondents participate in the current survey. The value of the parameter $\alpha_{i_1\ldots i_k}$ in the Dirichlet distribution can be set to $(m_{i_1\ldots i_k}/m)n$, where $m_{i_1\ldots i_k}$ 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 $\alpha_{i_1\ldots i_k}$ 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.

## 3. Testing approach

### 3.1. Multiple testing

In this section, we propose several multiple-testing methods for testing the relationship of $\pi_j$. Assume that there are $k$ responses and we are interested in testing

$$\left.\begin{array}{c} H_{01}:\pi_2 \leqslant \pi_1 \ \textit{versus}\ H_{11}:\pi_2 > \pi_1, \\ H_{02}:\pi_3 \leqslant \pi_2 \ \textit{versus}\ H_{12}:\pi_3 > \pi_2, \\ \vdots \\ H_{0k-1}:\pi_k \leqslant \pi_{k-1} \ \textit{versus}\ H_{1k-1}:\pi_k > \pi_{k-1}. \end{array}\right\} \tag{4}$$

It may be reasonable to test the hypothesis $\pi_1 = ... = \pi_k$ first, and then to proceed to test the one-sided test when the hypothesis $\pi_1 = ... = \pi_k$ is rejected. The approach for testing a point null hypothesis has been discussed by Berger (1985). It is necessary to assign a probability $\xi_0$ to the case $H_0 : \pi_1 = ... = \pi_k$ and to spread out the probability of $1 - \xi_0$ on the alternative hypothesis $H_0^c$. Since the probability of the fact $\pi_1 = ... = \pi_k$ may be low, we do not investigate testing the point null hypothesis in detail in this study. In addition, according to the ranking criterion (10) that is used in this study, for ranking two responses $\pi_i$ and $\pi_j$, both one-sided hypotheses $H_0 : \pi_i > \pi_j$ and $H_0 : \pi_j > \pi_i$ are considered, which may reflect the information that is obtained from the point null hypothesis.

For testing expressions (4), the decision rules that are considered here are to control the posterior expected false discovery rate. The concept of the false discovery rate was proposed by Benjamini and Hochberg (1995) to determine optimal thresholds for a multiple-testing setting. For testing multiple hypotheses, the possible outcomes (over the $l$ tests) may be summarized as in Table 1.

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 notation and definitions are given. Let $z_i$ denote an indicator that the $i$th hypothesis $H_{0i}$ is false and let $u_i = P(z_i = 1|\mathbf{n})$ denote the marginal posterior probability of $\pi_{i+1} > \pi_i$. The rejection of $H_{0i}$ is denoted by $d_i = 1$; otherwise $d_i = 0$. Let $z = (z_1, \ldots, z_{k-1})$ and $d = (d_1, \ldots, d_{k-1})$. Under the frequentist set-up, the false discovery rate and false negative rate are denoted by the expectations $E[V/(D+\varepsilon)]$ and $E[T/(n-D+\varepsilon)]$ respectively, where $D = \Sigma_{i=1}^{k-1} d_i$ and $\varepsilon$ is a small constant to avoid a zero denominator. In real applications, $\varepsilon$ can be chosen as 0.00001.

Let

$$\mathrm{FDR}(d, z) = \frac{\sum_{i=1}^{k-1} d_i(1 - z_i)}{D + \varepsilon}$$

**Table 1.** Outcomes of multiple tests†

| *Real state* | *Test result* | | *Number of hypotheses* |
|---|---|---|---|
| | *Number of $H_{0i}$ accepted* | *Number of $H_{0i}$ rejected* | |
| Number of true $H_{0i}$ | $U$ | $V$ | $l_0$ |
| Number of false $H_{0i}$ | $T$ | $S$ | $l_1$ |
| | $l - D$ | $D$ | $l$ |

†The notation $l$ is the total number of hypotheses, $l_0$ is the unknown number of true null hypotheses, $l_1$ is the unknown number of false null hypotheses, $V$ is the number of false positive results, $T$ is the number of false negative results, $S$ 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.

denote the false discovery rate and

$$\mathrm{FNR}(d,z)=\frac{\sum_{i=1}^{k-1}(1-d_i)z_i}{n-D+\varepsilon}$$

the false negative rate.

Under a Bayesian setting, these error rates are defined as the posterior expected false discovery rate denoted by $\overline{\mathrm{FDR}}(d,\mathbf{n})$ and the posterior expected false negative rate denoted as $\overline{\mathrm{FNR}}(d,\mathbf{n})$, where

$$\overline{\mathrm{FDR}}(d,n)=\frac{\sum_{i=1}^{k-1}d_i(1-u_i)}{D+\varepsilon}$$

and

$$\overline{\mathrm{FNR}}(d,\mathbf{n})=\frac{\sum_{i=1}^{k-1}(1-d_i)u_i}{n-D+\varepsilon}.$$

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

$$\overline{\mathrm{FD}}(d,\mathbf{n})=\sum_{i=1}^{k-1}d_i(1-u_i)$$

and

$$\overline{\mathrm{FN}}(d,\mathbf{n})=\sum_{i=1}^{k-1}(1-d_i)u_i.$$

### 3.2. Testing procedures

We here introduce several multiple-testing procedures from Berger (1985) and Muller *et al.* (2004) for testing expressions (4).

#### 3.2.1. Method 1

The decision to accept or reject the null hypothesis is based on the specific loss function that was proposed by Berger (1985), which is defined as

$$\left.\begin{array}{ll} 0 & \text{if the decision taken is right,} \\ c & \text{if we reject } H_{0i} \text{ when it is true,} \\ 1 & \text{if we accept } H_{0i} \text{ when it is false,} \end{array}\right\} \tag{5}$$

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

$$L_{\mathrm{N}}(d,\mathbf{n})=c\,\overline{\mathrm{FD}}+\overline{\mathrm{FN}}. \tag{6}$$

#### 3.2.2. Method 2

The second method is to consider the loss function

$$L_{\mathrm{R}}(d,\mathbf{n})=c\,\overline{\mathrm{FDR}}+\overline{\mathrm{FNR}}.$$

*3.2.3. Method 3*

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

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

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

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

$$d_i = I(u_i \geqslant t), \tag{7}$$

where $t$ are $t_{\mathrm{N}} = c/(c+1)$, $t_{\mathrm{R}}(\mathbf{n}) = u_{(n-D^*)}$ and $t_{2\mathrm{R}}(\mathbf{n}) = \min\{s : \mathrm{FDR}(s, \mathbf{n}) \leqslant e_{2\mathrm{R}}\}$ under the loss functions $L_{\mathrm{N}}$, $L_{\mathrm{R}}$ and $L_{2\mathrm{R}}$ respectively. In the expressions for $t_{\mathrm{R}}$ and $t_{2\mathrm{R}}$, $u_{(i)}$ is the $i$th order statistic of $\{u_1, \ldots, u_n\}$, and $D^*$ is the optimal number of discoveries 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.

The selections of $c$- and $e_{2\mathrm{R}}$-values in the loss functions may depend on the economic costs. In the real applications, if we think that the false positive rate is more serious than the false negative rate, then $c$ can be selected to be larger than 1 or $e_{2\mathrm{R}}$ is selected to be small. If we do not consider selecting $c$ and $e$ from the real application viewpoint, we can consider the problem in terms of the criterion of minimizing the penalty score that is proposed in Section 5.2. The related discussion is given in Section 5.2.

By definition, $u_i$ in the model for the multiple-response questions can be expressed as proportional to

$$\int \cdots \int \left\{ I\left(0 < \sum_{j=1}^{k} i_j \leqslant s\right) I(\pi_{l+1} > \pi_l) \frac{\Gamma\left(\sum\limits_{i_j = 0 \text{ or } 1} \alpha_{i_1 \ldots i_k} + n_{i_1 \ldots i_k}\right)}{\Gamma(\alpha_{i_1 \ldots i_k} + n_{i_1 \ldots i_k})} \prod_{i_j = 0 \text{ or } 1} p_{i_1 \ldots i_k}^{\alpha_{i_1 \ldots i_k} + n_{i_1 \ldots i_k}} \right\} \times \prod_{i_j = 0 \text{ or } 1} \mathrm{d}p_{i_1 \ldots i_k}, \tag{8}$$

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

*Theorem 1.* By normal approximation, the multiple integration (8) can be approximated by

$$\Phi(B/\sqrt{C}), \tag{9}$$

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

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

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

and

$$C = \frac{1}{A^2(A+1)} \left\{ \sum_{i_{l+1}=1, i_l = 0} (\alpha_{i_1 i_2 \ldots i_k} + n_{i_1 i_2 \ldots i_k})(A - \alpha_{i_1 i_2 \ldots i_k} - n_{i_1 i_2 \ldots i_k}) \right.$$

$$+\sum_{i_l=1,i_{l+1}=0}(\alpha_{i_1i_2\ldots i_k}+n_{i_1i_2\ldots i_k})(A-\alpha_{i_1i_2\ldots i_k}-n_{i_1i_2\ldots i_k})$$

$$\left.+2\sum_{i'_{l+1}=1,i'_l=0}\sum_{i''_l=1,i''_{l+1}=0}(\alpha_{i'_1i'_2\ldots i'_k}+n_{i'_1i'_2\ldots i'_k})(\alpha_{i''_1i''_2\ldots i''_k}+n_{i''_1i''_2\ldots i''_k})\right\}.$$

The proof is given in Appendix A.

To evaluate the performance of cumulative distribution function (9), we conduct a simulation to obtain the value of expression (8) and then compare it with the value of function (9). The results show that these two values are very close. However, it is much more time consuming to obtain expression (8) than to calculate function (9). Therefore, we conclude that function (9) is a more efficient formula to obtain $u_i$ than expression (8) is. The R code for calculating $u_i$ by function (9) and the data are available from

`http://www.blackwellpublishing.com/rss`

## 4. Ranking approach and ranking consistency

If not all of the hypotheses in expression (4) are rejected, there is not 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 in total $C_2^k$ hypotheses for $k$ responses. The rank of the $i$th response can be defined as

$$R_i = k - \sum_{j=1,j\neq i}^{k} I(\pi_i > \pi_j). \tag{10}$$

Using criterion (10), we define a response as 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 as 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.

### 4.1. Bayesian ranking consistency property

A test is called ranking consistent if $\pi_j = \pi_i$ is rejected by the test, and 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 $I_{\pi_j-\pi_g>0}$.

We provide Fig. 1 to illustrate the ranking consistency rule graphically. From the examples that were given in Wang (2008a), under the frequentist framework, the tests that are derived by the conventional approaches do not have the property of frequentist ranking consistency. It is still unknown whether there are 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 (7) that were considered in Section 3 for different $t$-values under the loss functions $L_{\mathrm{N}}$, $L_{\mathrm{R}}$ and $L_{2\mathrm{R}}$ are ranking consistent.

*Proof.* For the three tests in Section 3, the decision rules of the tests are based on decision (7). By this form, for a fixed cut-off $t$, the decision rule depends on only the Bayes estimator $u_i$ of $H_{0i}$. If a hypothesis $H_{0i}$ with a smaller $u_i$ is rejected, then a hypothesis $H_{0j}$ with a larger $u_j$ is accordingly rejected by the rule. Thus, the proof is complete.

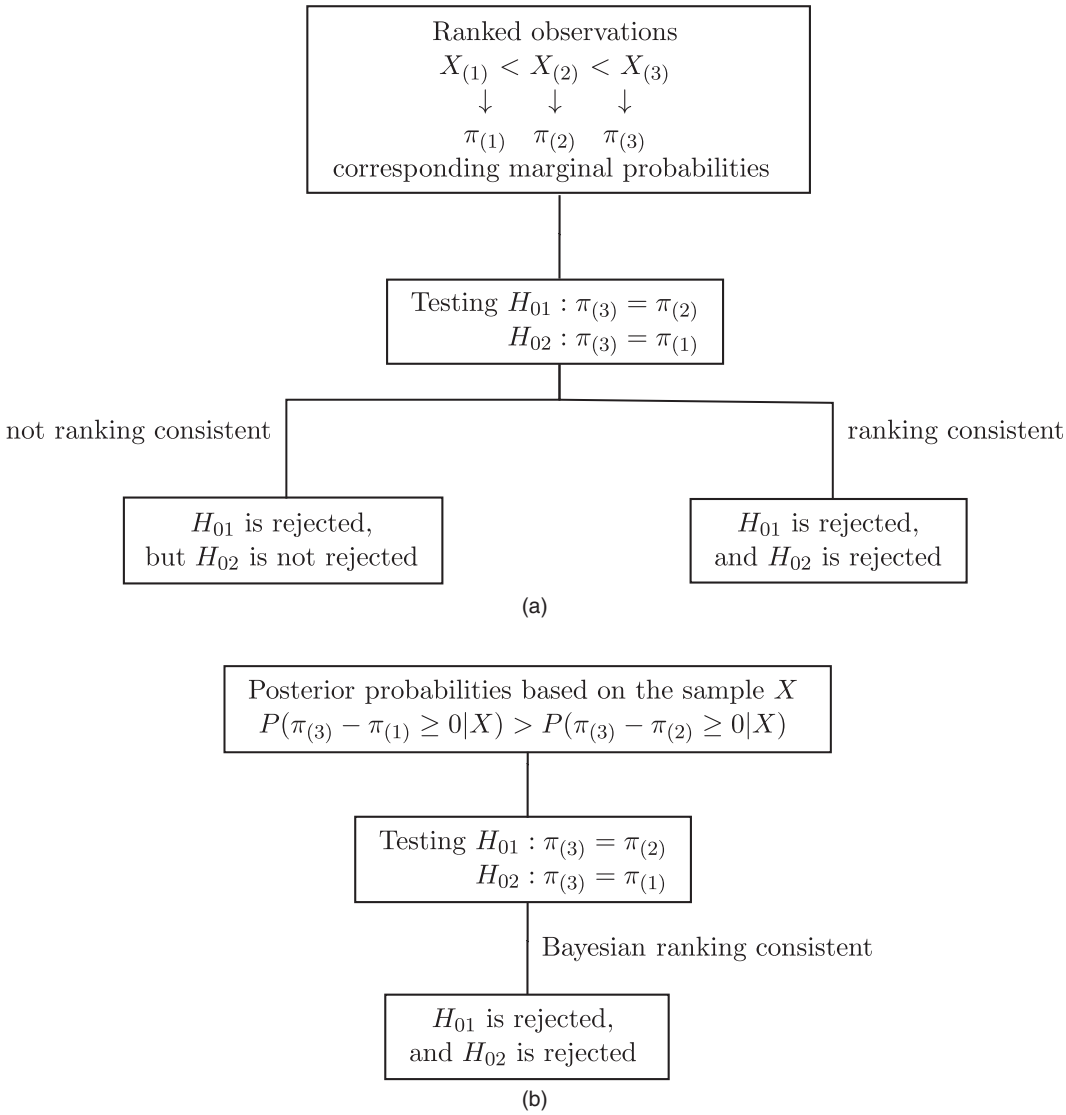


**Fig. 1.** (a) Frequentist and (b) Bayesian ranking consistency rule for ranking three responses

## 5. Simulation result

### *5.1. Rejection rate*

In this section, 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 $w_j = \Sigma_{i_j=1} \alpha_{i_1 \ldots i_k}$, $j = 1, \ldots, k$. The simulation procedure is first to generate a set of $p$ from the prior distribution, and then to use the $p$ to generate a set of $\mathbf{n}$. Next, calculate $u_i$ conditioning on the $\mathbf{n}$ for the three different loss criteria. To test the $k-1$ hypotheses of expressions (4), we can count the number of rejections for the $k-1$ hypotheses for the three methods. Although the probability of the validity 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 number of rejections for the hypothesis $H_{0i} : \pi_{i+1} < \pi_i$ of a good test should be close to $P(\pi_{i+1} > \pi_i)$. Thus, we can use this criterion to evaluate the testing methods. We repeat the simulation process 1000 times and the results are shown in Tables 2 and 3.

Besides the evaluation of the rejection rates of one of the four subhypotheses, we also provide the performance of the type I error and type II error of these methods for testing all the four null subhypotheses. The type I error is the probability that the rejection decision is reached when the generated $p$ is under hypothesis $H_0$ and the type II error is the probability that the rejection decision is not reached when the generated $p$ is under the alternative hypothesis $H_0^c$. We present the averages of type I error and type II error for the three methods in Table 4 and Table 5. The calculation of the average of the type I error is first to generate $p$, satisfying all of the four null subhypotheses, and to generate a set of $\mathbf{n}$ based on the $p$. We repeat the process 1000 times and calculate the average of the rejection rate for the 1000 replicates, which is the average of the type I error. The calculation of the average of type II error can be calculated similarly.

*5.1.1. Example 1*

Consider the case $k = 5$ 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 to 7. In this case, $w_1 = 133 = w_2 < w_3 = 147 < w_4 = 168 < w_5 = 203$. Under this set-up, 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 test expressions (4), we compare the three methods that were introduced in Section 3. The rejection rates for each method are listed in Table 2, where the $c$-values are 1 and 0.33 for $L_N$ and $L_R$

**Table 2.** Rejection rates of the three methods corresponding to each hypothesis in expressions (4) for 1000 replicates of example 1, where 'true probability' denotes the true probability of rejecting the null hypothesis under the prior distribution

| *Method* | *Rejection rates for the following hypotheses:* | | | |
|---|---|---|---|---|
| | $H_{01} : \pi_2 \leqslant \pi_1$ | $H_{02} : \pi_3 \leqslant \pi_2$ | $H_{03} : \pi_4 \leqslant \pi_3$ | $H_{04} : \pi_5 \leqslant \pi_4$ |
| True probability | 0.5 | 0.859 | 0.93 | 0.986 |
| $L_N$ | 0.529 | 0.942 | 0.981 | 0.999 |
| $L_R$ | 0.973 | 0.971 | 0.990 | 0.998 |
| $L_{2R}$ | 0.455 | 0.904 | 0.971 | 0.997 |

**Table 3.** Rejection rates of the three methods corresponding to each hypothesis in expressions (4) for 1000 replicates of example 2, where 'true probability' denotes the true probability of rejecting the null hypothesis under the prior distribution

| *Method* | *Rejection rates for the following hypotheses:* | | | |
|---|---|---|---|---|
| | $H_{01} : \pi_2 \leqslant \pi_1$ | $H_{02} : \pi_3 \leqslant \pi_2$ | $H_{03} : \pi_4 \leqslant \pi_3$ | $H_{04} : \pi_5 \leqslant \pi_4$ |
| True probability | 0.709 | 0.701 | 0.695 | 0.690 |
| $L_N$ | 0.790 | 0.771 | 0.764 | 0.781 |
| $L_R$ | 0.892 | 0.894 | 0.881 | 0.881 |
| $L_{2R}$ | 0.605 | 0.617 | 0.601 | 0.631 |

**Table 4.** Type I and type II errors as well as the rejection rates of the three methods for testing $H_0$: $\pi_2 \leqslant \pi_1, \pi_3 \leqslant \pi_2, \pi_4 \leqslant \pi_3, \pi_5 \leqslant \pi_4$ for 1000 replicates of example 1

| *Method* | *Rejection rate* | *Type I error under* $H_0$† | *Type II error under* $H_0^c$† |
|---|---|---|---|
| $L_N$ | 0.451 | 0.323 | 0 |
| $L_R$ | 0.925 | 0.938 | 0 |
| $L_{2R}$ | 0.350 | 0.236 | 0 |

†True probability of $H_0 = 0.347$.

**Table 5.** Type I and type II errors as well as the rejection rates of the three methods for testing $H_0$: $\pi_2 \leqslant \pi_1, \pi_3 \leqslant \pi_2, \pi_4 \leqslant \pi_3, \pi_5 \leqslant \pi_4$ for 1000 replicates of example 2

| *Method* | *Rejection rate* | *Type I error under* $H_0$† | *Type II error under* $H_0^0$† |
|---|---|---|---|
| $L_N$ | 0.254 | 0.262 | 0 |
| $L_R$ | 0.712 | 0.697 | 0 |
| $L_{2R}$ | 0.218 | 0.057 | 0.025 |

†True probability of $H_0 = 0.146$.

and the $e_{2R}$-value is 0.15 for $L_{2R}$. The selection of the $c$- and $e_{2R}$-values here is based on the evaluation of the penalty score, which is introduced in Section 5.2.

In Table 4, the average of the type II error is denoted 0, which is the result based on 1000 replicates. The true type II error should be very close to 0 but not exactly equal to 0.

*5.1.2. Example 2*

Consider the case $k = 5$ and a Dirichlet prior distribution on the parameter space with $\alpha_{00000} = 0$, $\alpha_{00001} = 56$, $\alpha_{00010} = 49$, $\alpha_{00100} = 42$, $\alpha_{01000} = 35$, $\alpha_{10000} = 28$ and the others are equal to 7. In this case, $w_1 = 133 < w_2 = 140 < w_3 = 147 < w_4 = 154 < w_5 = 161$. We have $P(\pi_2 > \pi_1) = 0.709$, $P(\pi_3 > \pi_2) = 0.701$, $P(\pi_4 > \pi_3) = 0.695$ and $P(\pi_5 > \pi_4) = 0.690$. To test expressions (4), we compare the three methods that were introduced in Section 3. The rejection rates for each method are listed in Table 3, where the $c$-values are 1 and 0.54 for $L_N$ and $L_R$ and the $e_{2R}$-value is 0.25 for $L_{2R}$.

The performances of the three methods for testing each subhypothesis are provided in Tables 2 and 3. 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. Overall, by comparing the rejection rates of the methods with the true probability of rejecting the hypothesis, method 1 and method 3 may be superior to method 2 in many cases, as shown in the simulation study.

From Tables 4 and 5, the performances of type I and II errors and the rejection rates of the three methods for testing all the four subhypotheses are similar to the performance of the rejection rates for testing one of the subhypotheses that are presented in Tables 2 and 3, which

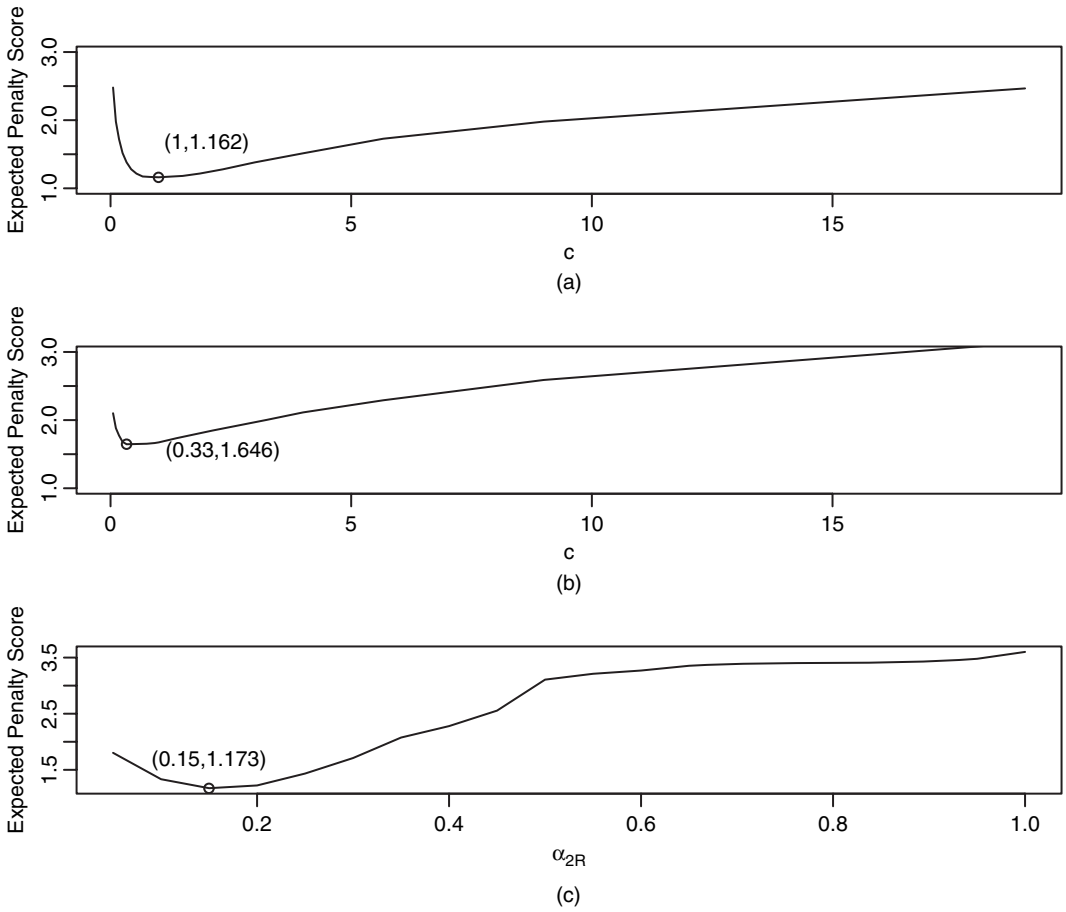


**Fig. 2.** Expected penalty scores of the three methods under the conditions of example 1: (a) loss function $L_N$; (b) loss function $L_R$; (c) loss function $L_{2R}$

show that method 1 and method 3 are superior to method 2. In this study, we adopt the critical rejection region that was suggested in Muller *et al.* (2004) for the three methods. Readers can adjust the critical region to reach a required type I error or type II error.

It is worth noting that we also conducted a simulation study to investigate the robustness of the three methods to the selection of the prior. In examples 1 and 2, the priors are assumed to be known. When the true prior is not known and the prior selected is not far from the true prior, the above comparison results of the three methods are similar to the known prior case. Thus, unless the prior selected is far from the true prior, the results of the three methods discussed above still hold.

### *5.2. Penalty score*

In this section, we shall set up a penalty score to evaluate the three methods in terms of ranking error. To rank the $i$th responses, for a given method, we need to calculate their $R_i$-values by using this method and then to use the $R_i$-value to rank the responses. A penalty score is defined as the summation of the absolute values of the difference between the true rank and the rank derived by the method. For example, in the case $k = 5$, if the true rank of the first response is 1, and the

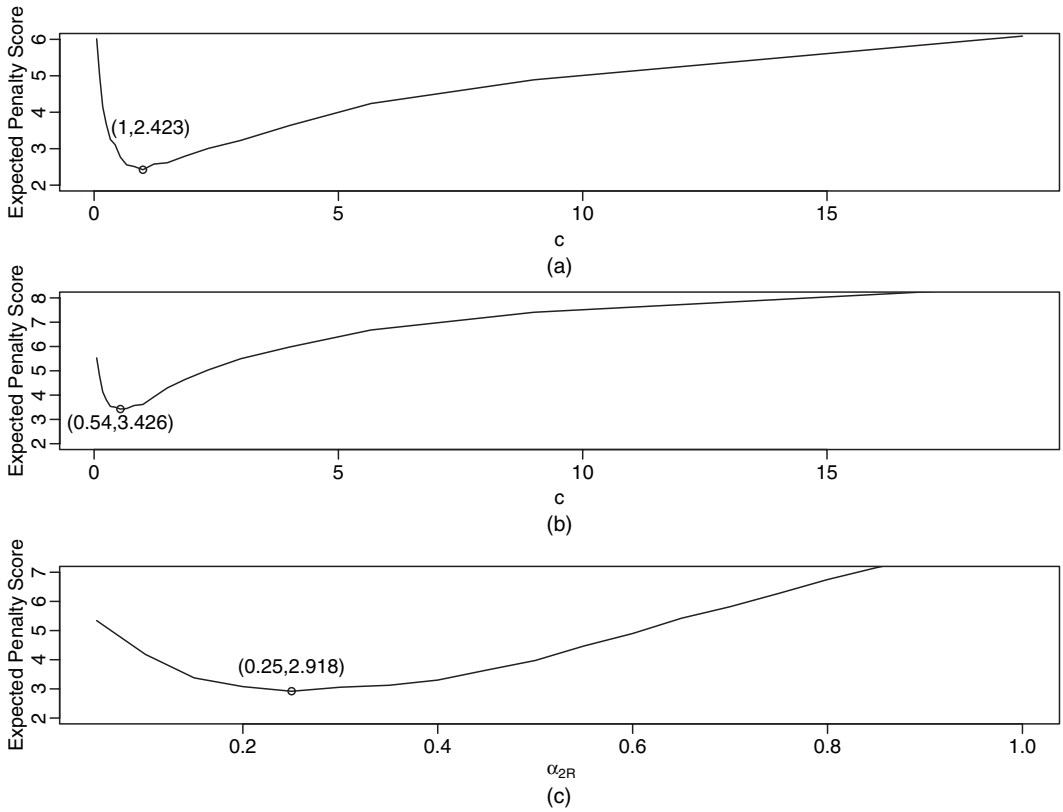


**Fig. 3.** Expected penalty scores of the three methods under the condition of example 2: (a) loss function $L_N$; (b) loss function $L_R$; (c) loss function $L_{2R}$

true rank of the second response is 2, etc. then we use the notation (1, 2, 3, 4, 5) to denote the true rank. If the rank that is 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*: set up a prior for $\alpha$.
*Step 2*: generate a set of $p$ from the prior distribution with respect to the $\alpha$-value in step 1. From this $p$, we can obtain a true rank for $\pi_i$ based on this $p$.
*Step 3*: using the probability mass function (1) with the $p$ in step 2, generate a set $\mathbf{n}$.
*Step 4*: set up the $\binom{k}{2}$ null hypotheses for any two different $\pi_i$. Then, on the basis of $\mathbf{n}$ in step 3, 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 expression (10) to rank the $k$ responses and calculate the penalty score from the rank derived and the true rank in step 2 for each method.
*Step 5*: repeat steps 2–4 1000 times. Take the average of the penalty scores in step 4 for each method and the approximated expected penalty score for each method is derived.

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 $e_{2R}$. In a real application, the selection of $c$ in methods 1 and 2 may depend on the true cost of an incorrect decision and the selection of $e_{2R}$ in method 3 may depend on the tolerance error allowed. However, from a theoretical viewpoint, we can investigate the situation of $c$ and $e_{2R}$ such that the three methods have the smallest penalty score.

Based on the simulation procedures, the performance of the expected penalty score for different $c$ and $e_{2R}$ corresponding to $\alpha_{i_1 \cdots i_k}$ in examples 1 and 2 are presented in Figs 2 and 3.

The minimum expected penalty scores for methods 1–3 are 1.162, 1.646 and 1.173 in Fig. 2, which occur at $c = 1$, $c = 0.33$ and $e_{2R} = 0.15$ respectively. The minimum expected penalty scores for methods 1–3 are 2.4323, 3.426 and 2.918 in Fig. 3, which occur at $c = 1$, $c = 0.54$ and $e_{2R} = 0.25$ respectively. Basically, Figs 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 lead to the worst performance of these three methods. From the viewpoint of ranking, this consequence coincides with the results in Section 5.1. The reject rates for the methods under the frequentist framework were presented in Wang (2008a).

Besides the priors that were used in examples 1 and 2, we conduct a simulation study to investigate the $c$- and $e_{2R}$-selection for other prior distributions as well. The simulation results show that the penalty scores are smaller when $c$ is selected to be close to 1 and 0.5 for the loss functions $L_N$ and $L_R$ respectively, and $e_{2R}$ is selected to be close to 0.2 for the loss function $L_{2R}$. Therefore, if there are no other preferences, the $c$ and $e_{2R}$ can be selected to be near the values suggested above.

## 6. Real data example

A real data example is used in this section to illustrate the methods and to present a case which is ranking inconsistent under the frequentist framework (Wang, 2008a) but is ranking consistent under the Bayesian framework. This survey example is of 49609 first-year college students in Taiwan regarding their preferences for college studies. The data set can be accessed at `http://www.stat.nctu.edu.tw/~hwang/ranking.htm` and is included in the on-line supplementary file. A multiple-response question from the questionnaire is shown as an example.

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

1. Read Chinese and foreign classics
2. Travel around Taiwan
3. Present academic papers in conferences
4. Lead large-scale activities
5. Be on a school team
6. Be a student association leader
7. Participate in internship programs
8. Fall in love
9. Have sexual experience
10. Travel around the world
11. Make many friends
12. Other'

Ranking the responses of this multiple-response question according to student preferences is of interest. To simplify the illustration, we do not consider the problem of ranking all the responses. Only these five responses are ranked: read Chinese and foreign classics, present academic papers in conferences, lead large-scale activities, be on a school team and be a student association leader.

The whole data set of this survey included 49609 interview data. These data can be used to obtain the true ranks of the five responses. To illustrate the methods, suppose that we do not have the whole data set but only the interview data of 100 randomly selected respondents. From the whole data set, the number of respondents who selected the five responses is 8858, 5358, 10578, 6823 and 12145. The first, second and third ranks indicate that the students prefer to be a student association leader, to lead large-scale activities and to read Chinese and foreign classics.

In this example, the notation $i_1 = 1$, $i_2 = 1$, $i_3 = 1$, $i_4 = 1$ and $i_5 = 1$ in $n_{i_1 i_2 i_3 i_4 i_5}$ corresponds to selection of the 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 leader respectively.

According to 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 $m_1 = 46$, $m_2 = 25$, $m_3 = 47$, $m_4 = 28$ and $m_5 = 58$. Let $m_{(i)}$ denote the corresponding order statistics of $m_i$. From the data, the most selected response is be a student association leader. Next is lead large-scale activities, followed by read Chinese and foreign classics. Consequently, we have $m_{(5)} = 58$, $m_{(4)} = 47$, $m_{(3)} = 46$, $m_{(2)} = 28$ and $m_{(1)} = 25$. Let $\pi_{(i)}$ denote the marginal probability corresponding to the order statistic $m_{(i)}$. Now we are interested in testing

$$\begin{aligned} &H_{01} : \pi_{(5)} \leqslant \pi_{(4)} \quad \textit{versus} \quad H_{11} : \pi_{(5)} > \pi_{(4)}, \\ &H_{02} : \pi_{(5)} \leqslant \pi_{(3)} \quad \textit{versus} \quad H_{12} : \pi_{(5)} > \pi_{(3)}. \end{aligned} \tag{11}$$

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 hypothesis $H_{01}$, 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 cut-off point of the standard normal distribution is 1.96, resulting in the rejection of $H_{01}$ by the two tests with type I error 0.025. However, when testing hypothesis $H_{02}$, the values of statistics corresponding to the Wald test and generalized score test are 1.59 and 1.57, which do not lead to the rejection of $H_{02}$ in either of the 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 set, 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.

To implement method 1 and method 2, we select $c = 1$ and $e_{2R} = 0.15$ corresponding to method 1 and method 2, resulting in $t = 0.5$ and $t = 0.01$ with respect to the two methods. For testing expression (11) under the given prior, we have $u_1 = 0.9949$ and $u_2 = 0.9922$. Consequently, by expression (7), $H_{01}$ and $H_{02}$ are both rejected by the two methods.

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

According to the simulation result, there is a proportion around 0.15 that the frequentist ranking inconsistency phenomenon occurs. The theoretical investigation of the occurrence of ranking inconsistency is still understudied. We also provide a frequentist ranking consistency sample for this example as follows.

We have 100 randomly selected data with $n_{10000}=14, n_{01000}=3, n_{00100}=4, n_{00010}=6, n_{00001}=13$, $n_{11000}=4$, $n_{10100}=2$, $n_{10010}=0$, $n_{10001}=7$, $n_{01100}=0$, $n_{01010}=0$, $n_{01001}=4$, $n_{00110}=1$, $n_{00101}=13, n_{00011}=2, n_{11100}=3, n_{11010}=1, n_{11001}=1, n_{01110}=0, n_{01101}=1, n_{01011}=0, n_{00111}=4$, $n_{10110}=1$, $n_{10101}=1$, $n_{10011}=3$, $n_{11110}=0$, $n_{11101}=0$, $n_{11011}=1$, $n_{10111}=3$, $n_{01111}=5$ and $n_{11111}=3$ for the 100 data. These lead to $m_1=44$, $m_2=26$, $m_3=41$, $m_4=30$ and $m_5=61$. Consequently, we have $m_{(5)}=61$, $m_{(4)}=44$, $m_{(3)}=41$, $m_{(2)}=30$ and $m_{(1)}=26$.

To test

$$\begin{aligned} &H_{01}:\pi_{(5)}\leqslant\pi_{(4)} \ \textit{versus} \ H_{11}:\pi_{(5)}>\pi_{(4)},\\ &H_{02}:\pi_{(5)}\leqslant\pi_{(3)} \ \textit{versus} \ H_{12}:\pi_{(5)}>\pi_{(3)}, \end{aligned} \tag{12}$$

the two statistics values for the Wald test and generalized score test under the frequentist framework are 2.12 and 2.08, resulting in the rejection of hypothesis $H_{01}$ by the two tests with type I error 0.025. When testing hypothesis $H_{02}$, the values of statistics corresponding to the Wald test and generalized score test are 3.24 and 3.09, which also lead to the rejection of $H_{02}$ in either of the two tests. The above sample leads to ranking consistency for the Wald and score tests under the frequentist framework.

## 7. Conclusions

This study establishes Bayesian multiple-testing procedures under the false discovery rate and loss functions criteria for investigating the ranking of responses in a multiple-response question. The test statistic is based on the posterior probability, and an approximate formula for the posterior probability is provided.

The simulation study indicates that the use of the loss functions $L_{\mathrm{N}}$ and $L_{2\mathrm{R}}$ is better than that of the loss function $L_{\mathrm{R}}$ if we consider cases where $c$ and $e_{2\mathrm{R}}$ are selected so that the minimum expected penalty score occurs. However, in real applications, selection of the constant $c$ in $L_{\mathrm{N}}$ and $L_{\mathrm{R}}$ may depend on empirical information or economic costs, which may be determined by an experienced manager. The same is true for the selection of $e_{2\mathrm{R}}$. The set-up of $e_{2\mathrm{R}}$ may depend on the allowed tolerance error in real applications because $e_{2\mathrm{R}}$ provides a tolerance error for the false discovery rate.

This study also proposes an approach for ranking the responses of multiple-response questions under the Bayesian framework based on multiple-testing procedures. Conventional tests under the frequentist set-up do not have the ranking consistency property. Compared with methods under the frequentist framework, this Bayesian approach provides more convincing results because it has the Bayesian ranking consistency property.

This method can be applied to numerous other applications such as medical, social and psychological studies. Another important issue for analysing the multiple-response question is regarding the correlations between a multiple-response question and a single-response question or between two multiple-response questions. Future studies may propose approaches for exploring the association between two questions under the Bayesian framework.

## Acknowledgements

The authors thank the referees for the valuable comments that have substantially improved

the paper and Shao-Yuan Chang for computational assistance in this study. This work was partially supported by the National Science Council, National Center for Theoretical Sciences and 'Aiming for the top university program' of the National Chiao Tung University and Ministry of Education, Taiwan, Republic of China.

## Appendix A: Proof of theorem 1

In this proof, we derive the normal approximation formula for the multiple integration.

For a given $\mathbf{n}$, let $A = \Sigma_{i_j=0 \text{ or } 1}(\alpha_{i_1 i_2 \ldots i_k} + n_{i_1 i_2 \ldots i_k})$. From the property of the Dirichlet distribution, we have the expectation and variance of $p_{i'_1 i'_2 \ldots i'_k}$ equal to

$$\frac{\alpha_{i'_1 i'_2 \ldots i'_k} + n_{i'_1 i'_2 \ldots i'_k}}{A}$$

and

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

respectively.

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

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

Note that $\pi_{l+1} - \pi_l = \pi_{l+1} - \pi_{(l+1)l} - (\pi_l - \pi_{(l+1)l})$, where $\pi_{jl} = \Sigma_{i_j = i_l = 1} p_{i_1 \ldots i_k}$ denotes the corresponding probability that both the $j$th and the $l$th responses are selected.

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

$$\begin{aligned} \operatorname{var}(\pi_{l+1} - \pi_l) &= \operatorname{var}\{(\pi_{l+1} - \pi_{(l+1)l}) - (\pi_l - \pi_{(l+1)l})\} \\ &= \operatorname{var}(\pi_{l+1} - \pi_{(l+1)l}) + \operatorname{var}(\pi_l - \pi_{(l+1)l}) - 2\operatorname{cov}\{(\pi_{l+1} - \pi_{(l+1)l})(\pi_l - \pi_{(l+1)l})\} \\ &= C. \end{aligned} \tag{13}$$

Since the Dirichlet distribution belongs to the multiparameter exponential family (Bickel and Doksum, 2007), by the property of the exponential family, we can apply the central limit theory to obtain the normal approximation,

$$\pi_{l+1} - \pi_l \sim N(B, \sqrt{C}).$$

Thus, we have

$$\begin{aligned} v_l &= P(\pi_{l+1} - \pi_l > 0) \\ &= P\left\{ \frac{\pi_{l+1} - \pi_l - E[\pi_{l+1} - \pi_l]}{\sqrt{\operatorname{var}(\pi_{l+1} - \pi_l)}} > \frac{-E[\pi_{l+1} - \pi_l]}{\sqrt{\operatorname{var}(\pi_{l+1} - \pi_l)}} \right\} \\ &= P\left\{ Z > \frac{-E[\pi_{l+1} - \pi_l]}{\sqrt{\operatorname{var}(\pi_{l+1} - \pi_l)}} \right\} \\ &= \Phi\left\{ \frac{E[\pi_{l+1} - \pi_l]}{\sqrt{\operatorname{var}(\pi_{l+1} - \pi_l)}} \right\} \end{aligned} \tag{14}$$

The proof is complete.

## References

Agresti, A. and Liu, I. M. (1999) Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics*, **55**, 936–943.

Agresti, A. and Liu, I. M. (2001) Strategies for modeling a categorical variable allowing multiple category choices. *Sociol. Meth. Res.*, **29**, 403–434.
Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rates: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.
Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.
Berger, J. O. and Deely, J. (1988) A Bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *J. Am. Statist. Ass.*, **83**, 364–373.
Bickel, P. J. and Doksum, K. A. (2007) *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. 1, 2nd edn. Upper Saddle River: Pearson Prentice Hall.
Bilder, C. R., Loughin, T. M. and Nettleton, D. (2000) Multiple marginal independence testing for pick any/c variables. *Communs Statist. Simuln Computn*, **29**, 1285–1316.
Decady, Y. J. and Thomas, D. H. (2000) A simple test of association for contingency tables with multiple column responses. *Biometrics*, **56**, 893–896.
Do, K. A., Müller, P. and Tang, F. (2005) A Bayesian mixture model for differential gene expression. *Appl. Statist.*, **54**, 627–644.
Gonen, M., Westfall, P. H. and Johnson, W. O. (2003) Bayesian multiple testing for two-sample multivariate endpoints. *Biometrics*, **59**, 76–82.
Gopalan, R. and Berry, D. A. (1998) Bayesian multiple comparisons using Dirichlet process priors. *J. Am. Statist. Ass.*, **93**, 1130–1139.
Jäckle, A. and Lynn, P. (2007) Dependent interviewing and seam effects in work history data. *J. Off. Statist.*, **23**, 529–552.
Jenkins, S. P., Cappellari, L., Lynn, P., Jackle, A. and Sala, E. (2006) Patterns of consent: evidence from a general household survey. *J. R. Statist. Soc.* A, **169**, 701–722.
Lin, R., Louis, T. A., Paddock, S. M. and Ridgeway, G. (2006) Loss function based ranking in two-stage hierarchical models. *Baysn Anal.*, **1**, 915–946.
Loughin, T. M. and Scherer, P. N. (1998) Testing for association in contingency tables with multiple column responses. *Biometrics*, **54**, 630–637.
Lynn, P., Jäckle, A., Jenkins, S. P. and Sala, E. (2012) The impact of questioning method on measurement error in panel survey measures of benefit receipt: evidence from a validation study. *J. R. Statist. Soc.* A, **175**, 289–308.
Miranda-Moreno, L. F., Labbe, A. and Fu, L. (2007) Bayesian multiple testing procedures for hotspot identification. *Accid. Anal. Prevn*, **39**, 1192–1201.
Muller, P., Parmigiani, G. and Rice, K. (2007) FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8* (eds J. M. Bernardo, S. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West). Oxford: Oxford University Press.
Muller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Am. Statist. Ass.*, **99**, 990–1001.
Pammer, S., Fong, D. K. H. and Arnold, S. F. (2000) Forecasting the penetration of a new product: a Bayesian approach. *J. Bus. Econ. Statist.*, **18**, 428–435.
Scott, J. (2009) Nonparametric Bayesian multiple testing for longitudinal performance stratification. *Ann. Appl. Statist.*, **3**, 1655–1674.
Scott, J. G. and Berger, J. O. (2006) An exploration of aspects of Bayesian multiple testing. *J. Statist. Planng Inf.*, **136**, 2144–2162.
Thissen, D. and Steinberg, L. (1984) A response model for multiple choice items. *Psychometrika*, **49**, 501–519.
Thissen, D. and Steinberg, L. (1986) A taxonomy of item response models. *Psychometrika*, **51**, 567–577.
Umesh, U. N. (1995) Predicting nominal variable relationships with multiple responses. *J. Forecast.*, **14**, 585–596.
Wang, H. (2008a) Ranking responses in multiple choice questions. *J. Appl. Statist.*, **35**, 465–474.
Wang, H. (2008b) Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *J. Multiv. Anal.*, **99**, 896–911.