

Decision Theoretic Bayesian Hypothesis Testing with the Selection Goal

Naveen K. Bansal

Department of Mathematics, Statistics, and Computer Science,

P.O. Box 1881,

Marquette University, Milwaukee, WI 53201, U.S.A.

naveen.bansal@marquette.edu

Abstract: Consider a probability model $P_{\theta, \alpha}$, where $\theta = (\theta_1, \theta_2, \dots, \theta_k)^T$ is a parameter vector of interest, and α is some nuisance parameter. The problem of testing null hypothesis $H_0 : \theta_1 = \theta_2 = \dots = \theta_k$ against selecting one of k alternative hypotheses $H_i : \theta_i = \theta_{[k]} > \theta_{[1]}$, $i = 1, 2, \dots, k$, where $\theta_{[k]} = \max\{\theta_1, \theta_2, \dots, \theta_k\}$ and $\theta_{[1]} = \min\{\theta_1, \theta_2, \dots, \theta_k\}$, is formulated from a Bayesian decision theoretic point of view. This problem can be viewed as selecting a component with the largest parameter value if the null hypothesis is rejected. General results are obtained for the Bayes rule under monotonic permutation invariant loss functions. Bayes rules are obtained for k one-parameter exponential families of distributions under conjugate priors. The example of normal populations is considered in more detail under the non-informative (improper) priors. It is demonstrated through this example that the classical hypothesis testing yields a poor power as compared to the Bayes rules when the alternatives are such that a small fraction of the components of θ have significantly high values while most of them have low values. Consequences of this for the high dimensional data such as microarray data are pointed out.

MSC: 62F07; 62F15; 62F03; 62C10

Keywords: Selection of populations; Exponential families; Bayes decision rule; Bayes factor; Microarray; Multiple comparisons; Decreasing in Transposition property.

1 Introduction

Consider a problem of finding the best treatment from k given treatments. A typical statistical practice is to first test the null hypothesis of equal treatment effects. If the

null hypothesis is rejected, then post hoc tests such as Tukey’s multiple comparison test are performed to compare all or some pairs of treatments. Such procedures are ad hoc and unsatisfactory when the objective is to find the best treatment as pointed out by Berger and Deely (1988). The drawbacks arise in both hypothesis testing and in multiple comparisons. It is clear that the classical hypothesis tests such as the classical ANOVA will usually result in low power when the alternatives are such that only a small fraction of the treatments have high values while most of them have low values. This follows from the fact that they are uniformly best rotation invariant tests and distribute the power uniformly in all directions. To illustrate this point, consider, for example, three normal populations with different means but equal variances. The sample point with sample means $(\bar{y}_1, \bar{y}_2, \bar{y}_3) = (1.0, 0.2, 0.0)$ should yield a better evidence for selecting the first population as the best than the sample point with sample means $(\bar{y}_1, \bar{y}_2, \bar{y}_3) = (1.0, 0.8, 0.0)$. However, the test statistic such as F-statistic gives equal preference to both the points. In a high dimensional case, this problem becomes even more severe. To illustrate this, consider the number of independent components $k = 100$, and the sample size $n = 20$. Under the assumption of normal distribution with known variance, say $\sigma^2 = 1$, the likelihood ratio test statistic for testing $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ is given by $\chi^2 = 20 \sum_{i=1}^{100} (\bar{Y}_i - \bar{\bar{Y}})^2$. Now suppose, the observed data yields sample means with lots of zeros or near zeros, say, $(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_{100}) = (0.1, 0.02, 0.0, \dots, 0.0)$, then the p -value is almost 1 suggesting that the null hypothesis should be accepted; although, intuitively, it looks that there should be some evidence in support of the first component as the best.

The drawback in the multiple comparisons occurs due to the controlled Family Wise Error Rate (FWER), especially for the high dimensional data such as microarray data (Efron, et. al., 2001; Tusher, et. al., 2001; Ishwaran and Rao, 2003). Since too many comparisons are made while keeping the FWER controlled, the power of detecting a difference, and thus of selecting the best component, becomes very low (Benjamini and Hochberg, 1995; Westfall and Young, 1993). Although, an alternative method based on the False Discovery Rate (FDR) has been proposed by Benjamini and Hochberg (1995) (see also Benjamini and Yekutieli, 2005), it only controls the number of false rejections, and does not address the problem of selecting the best component.

In this paper, we propose a different approach by testing the null hypothesis against selecting one of k alternative hypotheses, where the i^{th} alternative hypothesis

states that the i^{th} component is the best. From the decision theoretic point of view, this problem can be defined as taking action from the action space $\mathcal{A} = \{0, 1, \dots, k\}$, where the action "0" means that the null hypothesis is accepted, and the action " i " ($i = 1, 2, \dots, k$) means that the i^{th} component is selected as the best. We develop a Bayesian decision theoretic methodology for this problem along the lines of Bayesian hypothesis testing.

The approach discussed in this paper may be useful for the high dimensional data such as cDNA microarray data. A typical cDNA microarray data consists of fluorescence intensities corresponding to the thousands of genes describing the expression levels of the genes in a mixture of DNA samples of two types of cells (e.g., cancer cells and normal cells). The higher intensity levels corresponding to a gene signals that the gene is responsible for the protein synthesis in the unhealthy cell. It is known that only few of the genes corresponds to the significant level of fluorescence intensities (Efron, et. al., 2001). Most of the statistical analyses in the literature use different variants of multiple comparisons to find these genes (Tusher, et. al., 2001). However, due to high dimensionality and small sample sizes, these methods do not yield significant results, which may be perhaps due to the fact that the acceptance regions of the standard hypothesis tests are too big. We believe that the statistical framework of testing the null hypothesis against selecting the best discussed in this paper may perhaps minimize this problem. Although an objective of the microarray data analysis is to detect several highly expressed genes and not just the most expressive gene, we will explain how the method discussed in this paper can be utilized to detect several genes.

The main idea of the problem discussed in this paper come from the decision theoretic approach to *Ranking and Selection* problems. The selection problem of selecting the best component has been considered quite extensively in the statistical literature. However, most of these works do not deal with the null hypothesis; in other word, they only consider the action space $\{1, 2, \dots, k\}$ without considering the possibility that the null hypothesis may be accepted. There are basically three approaches to the selection problem, the *indifference-zone approach* (Bechhofer, 1954), the *subset selection approach* (Gupta, 1956), and the decision theoretic approach (Bahadur, 1950; Bahadur and Goodman, 1952). While there is no mechanism in the *indifference-zone approach* or in the *subset selection approach* for including the null hypothesis, in the decision theoretic approach, the null hypothesis can easily be implemented as we will

see in the Section 2. Few authors have considered the null hypothesis with the selection goal. Karlin and Truax (1960) considered a slippage type of alternative and proved some optimal result for the symmetric decision rules. Berger and Deely (1988) gave the Bayesian formulation for the normal models using a hierarchical prior. The present work is in the same spirit of Berger and Deely's work, but we provide a general framework with the general loss, and with the general prior settings of Bayesian hypothesis testing.

The article is organized as follows. In Section 2, we present a general framework of the problem and derive Bayes rules for general loss functions. In Section 3, we consider some specific loss functions and obtain the corresponding Bayes rules. In Section 4, we implement the results of Section 3 for k independent one-parameter exponential families of distributions. In section 5, we consider normal populations and demonstrate how the current approach present a better alternative to the classical approach of hypothesis testing. We end with concluding remarks in Section 6.

2 General Formulation

Consider a probability model $P_{\boldsymbol{\theta}, \alpha}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)^T$ is a parameter vector of interest belonging to a parameter space $\Theta \subseteq \mathbb{R}^k$, and α is a nuisance parameter belonging to a space Ω . We assume that the space $\Theta_0 = \{\boldsymbol{\theta} \in \Theta : \theta_1 = \theta_2 = \dots = \theta_k\}$ is a subset of the parameter space Θ . A typical hypothesis testing problem tests the null hypothesis $H_0 : \theta_1 = \theta_2 = \dots = \theta_k$ against the alternative $H_a : \theta_i \neq \theta_j$ for some $i \neq j$. However, suppose the main interest is to select the component of $\boldsymbol{\theta}$ that is associated with the largest $\theta_{[k]} = \max\{\theta_1, \dots, \theta_k\}$, provided the null hypothesis is not true. A decision theoretic formulation of this can be given as follows: Let the action space be $\mathcal{A} = \{0, 1, \dots, k\}$, where the action "0" means that the null hypothesis is accepted, and the action " i " ($i = 1, \dots, k$) means that the hypothesis $H_i : \theta_i = \theta_{[k]} > \theta_{[1]}$ ($\theta_{[1]} = \min\{\theta_1, \dots, \theta_k\}$) is selected, in other words, the i^{th} component of $\boldsymbol{\theta}$ is selected as the best, i.e., the largest. This problem can be rephrased as testing null hypothesis $H_0 : \theta_1 = \theta_2 = \dots = \theta_k$ against selecting one of the k alternatives $H_i : \theta_i = \theta_{[k]} > \theta_{[1]}$, $i = 1, \dots, k$.

Let $f(\mathbf{y}|\boldsymbol{\theta}, \alpha)$ denote the density of $P_{\boldsymbol{\theta}, \alpha}$ with respect to a σ -finite measure μ at an observed random sample \mathbf{y} , and let Π denote the prior on Θ with the generic random variable denoted by $\boldsymbol{\Xi} = (\Xi_1, \dots, \Xi_k)^T$. If $L(\boldsymbol{\theta}, i)$, $i = 0, 1, \dots, k$, denotes the

loss for taking action i when $\Xi = \boldsymbol{\theta}$, then the Bayes risk of a randomized rule $\boldsymbol{\phi}(\mathbf{y}) = (\phi_0(\mathbf{y}), \phi_1(\mathbf{y}), \dots, \phi_k(\mathbf{y}))$ ($\sum_{i=0}^k \phi_i(\mathbf{y}) = 1$, and $0 \leq \phi_i(\mathbf{y}) \leq 1$) is given by

$$\begin{aligned} r(\Pi, \boldsymbol{\phi}) &= \sum_{i=0}^k EL(\Xi, i)\phi_i(\mathbf{Y}) \\ &= \sum_{i=0}^k E[\phi_i(\mathbf{Y})E(L(\Xi, i)|\mathbf{Y})]. \end{aligned}$$

A Bayes rule that minimizes the above Bayes risk is thus given by

$$\phi_j^B(\mathbf{y}) = \begin{cases} |N(\mathbf{y})|^{-1} & \text{if } j \in N(\mathbf{y}) \\ 0 & \text{if } j \notin N(\mathbf{y}) \end{cases} \quad (1)$$

where

$$N(\mathbf{y}) = \{i : E[L(\Xi, i)|\mathbf{y}] = \min_{j=0,1,\dots,k} E[L(\Xi, j)|\mathbf{y}]\}. \quad (2)$$

2.1 Prior Distribution and the Bayes rule

Consider the following hierarchical structure of the prior along the lines of Bayesian hypothesis testing (Berger, 1985).

$$\begin{aligned} \pi(H_0) &= p, & \pi(\boldsymbol{\theta} | H_0, \alpha) &= g_0(\boldsymbol{\theta}|\alpha) \\ \pi(H_a) &= 1 - p, & \pi(\boldsymbol{\theta} | H_a, \alpha) &= g_a(\boldsymbol{\theta}|\alpha) \end{aligned} \quad (3)$$

where g_0 and g_a are the densities with respect to a σ -finite measure ν such that $g_0(\boldsymbol{\theta}|\alpha) = 0$ for $\boldsymbol{\theta} \in \Theta \setminus \Theta_0$ and $g_a(\boldsymbol{\theta}|\alpha) = 0$ for $\boldsymbol{\theta} \in \Theta_0$, for all α . In other words, the conditional prior on Θ given α has ν -density $pg_0(\boldsymbol{\theta}|\alpha) + (1-p)g_a(\boldsymbol{\theta}|\alpha)$, where g_0 and g_a are the conditional densities with supports in Θ_0 and $\Theta \setminus \Theta_0$, respectively. Let the prior on Ω be given by the density $\omega(\alpha)$ with respect to a σ -finite measure ν_1 .

Let $\pi(\boldsymbol{\theta} | \mathbf{y}, H_0)$ denote the posterior ν -density with support in Θ_0 with respect to the prior density $g_0(\boldsymbol{\theta}|\alpha)\omega(\alpha)$, and let $m_0(\mathbf{y})$ be the corresponding marginal density of \mathbf{Y} . Let $\pi(\boldsymbol{\theta} | \mathbf{y}, H_a)$ denote the posterior ν -density with support in $\Theta \setminus \Theta_0$ with respect to the prior density $g_a(\boldsymbol{\theta}|\alpha)\omega(\alpha)$, and let $m_a(\mathbf{y})$ be the corresponding marginal density of \mathbf{Y} . Then, it can be seen, from (3), that the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Y} = \mathbf{y}$ is given by

$$\pi(\boldsymbol{\theta} | \mathbf{y}) = \pi(H_0 | \mathbf{y}) \pi(\boldsymbol{\theta} | \mathbf{y}, H_0) + \pi(H_a | \mathbf{y}) \pi(\boldsymbol{\theta} | \mathbf{y}, H_a), \quad (4)$$

where

$$\pi(H_0 | \mathbf{y}) = \frac{p m_0(\mathbf{y})}{p m_0(\mathbf{y}) + (1-p) m_a(\mathbf{y})} = p_0(\mathbf{y}), \text{ say} \quad (5)$$

$$\pi(H_a | \mathbf{y}) = \frac{(1-p) m_a(\mathbf{y})}{p m_0(\mathbf{y}) + (1-p) m_a(\mathbf{y})} = 1 - p_0(\mathbf{y}). \quad (6)$$

From (4), for $j = 0, 1, \dots, k$,

$$\begin{aligned} E[L(\Xi, j) | \mathbf{y}] &= p_0(\mathbf{y}) \int_{\Theta_0} L(\boldsymbol{\theta}, j) \pi(\boldsymbol{\theta} | \mathbf{y}, H_0) d\nu(\boldsymbol{\theta}) \\ &\quad + (1 - p_0(\mathbf{y})) \int_{\Theta} L(\boldsymbol{\theta}, j) \pi(\boldsymbol{\theta} | \mathbf{y}, H_a) d\nu(\boldsymbol{\theta}) \end{aligned} \quad (7)$$

Now assume that the loss for selecting j ($j = 0, 1, \dots, k$) at $\boldsymbol{\theta} \in \Theta_0$ is constant, i.e., at $\boldsymbol{\theta} \in \Theta_0$, $L(\boldsymbol{\theta}, 0) = l_0$, and for $j = 1, 2, \dots, k$, $L(\boldsymbol{\theta}, j) = l_1$, where $l_0 < l_1$. Note that the requirement $l_0 < l_1$ is imposed since the loss for selecting $j = 0$ must be smaller than the loss for selecting $j = 1, 2, \dots, k$ when $\boldsymbol{\theta} \in \Theta_0$.

Now, from (1) and (2), the Bayes rule accepts H_0 , i.e., selects $j = 0$, if for all $j = 1, 2, \dots, k$, $E[L(\Xi, 0) | \mathbf{y}] \leq E[L(\Xi, j) | \mathbf{y}]$ and thus from (7), accepts H_0 if for all $j = 1, 2, \dots, k$,

$$\int_{\Theta} [L(\boldsymbol{\theta}, 0) - L(\boldsymbol{\theta}, j)] \pi(\boldsymbol{\theta} | \mathbf{y}, H_a) d\nu(\boldsymbol{\theta}) \leq \frac{p}{1-p} B(\mathbf{y})(l_1 - l_0), \quad (8)$$

where $B(\mathbf{y}) = [p_0(\mathbf{y})/(1 - p_0(\mathbf{y}))]/[p/(1 - p)]$ is the Bayes factor.

If (8) does not hold for some $j = 1, 2, \dots, k$, then H_0 is rejected, and in that case, from (1) and (2), H_j is selected according to the smallest of $E[L(\Xi, i) | \mathbf{y}]$, $i = 1, 2, \dots, k$. And since $L(\boldsymbol{\theta}, j) = l_1$ for all $j = 1, 2, \dots, k$ when $\boldsymbol{\theta} \in \Theta_0$, from (7), this is equivalent to the smallest of

$$E[L(\Xi, j) | \mathbf{y}, H_a] = \int_{\Theta} L(\boldsymbol{\theta}, j) \pi(\boldsymbol{\theta} | \mathbf{y}, H_a) d\nu(\boldsymbol{\theta}). \quad (9)$$

Thus we have the following result.

Theorem 1 *Under the prior (3), the Bayes decision rule is given by*

Accept the null hypothesis $H_0 : \theta_1 = \theta_2 = \dots = \theta_k$, if for all $i = 1, \dots, k$,

$$E[(L(\Xi, 0) - L(\Xi, i)) | \mathbf{y}, H_a] \leq \frac{p}{1-p} B(\mathbf{y})(l_1 - l_0). \quad (10)$$

Otherwise, select $H_j : \theta_j = \theta_{[k]} > \theta_{[1]}$ for which

$$E[L(\Xi, j) | \mathbf{y}, H_a] = \min_{i=1,2,\dots,k} E[L(\Xi, i) | \mathbf{y}, H_a]. \quad (11)$$

If the equality above is attained at more than one j then select one of the associated H_j with equal probability.

Remark 1 Under the above methodology, only one component is selected upon rejection of H_0 . However, the above methodology can be applied to the situations where it may be desirable to select more than one component; for example, in microarray data analysis, where it is generally desirable to select more than one highly expressed genes. This can be done by selecting all components i for which the reverse of the inequality in (10) hold.

In order to obtain a more specific form of the Bayes decision rule, we now put the following restrictions on the loss functions that are based on the monotonicity properties of the loss, c.f., Eaton(1967).

Assumption A For $\theta \in \Theta \setminus \Theta_0$,

- (i) $L(\rho(\boldsymbol{\theta}), 0) = L(\boldsymbol{\theta}, 0)$ for all permutations ρ .
- (ii) $L(\rho(\boldsymbol{\theta}), i) = L(\boldsymbol{\theta}, \rho(i))$, for all permutations ρ .
- (iii) $L(\boldsymbol{\theta}, i_1) \leq L(\boldsymbol{\theta}, i_2)$ if $\theta_{i_1} \geq \theta_{i_2}$ for all $i_1 \neq i_2$.

In the above, $\rho(\boldsymbol{\theta}) = (\theta_{\rho(1)}, \theta_{\rho(2)}, \dots, \theta_{\rho(k)})^T$, and $(\rho(1), \rho(2), \dots, \rho(k))$ is a permutation of $(1, 2, \dots, k)$. We shall also assume that $\rho(\Theta) = \Theta$ for all permutations ρ .

Now suppose the density $f(\mathbf{y}; \boldsymbol{\theta}, \alpha)$ can be reduced to the following form:

$$f(\mathbf{y}; \boldsymbol{\theta}, \alpha) = h(\mathbf{t}(\mathbf{y}), \mathbf{s}(\mathbf{y}); \boldsymbol{\theta}, \alpha) r(\mathbf{y}), \quad (12)$$

where $r(\mathbf{y})$ is independent of $\boldsymbol{\theta}$ and α , and $(\mathbf{t}(\mathbf{y}), \mathbf{s}(\mathbf{y}))$ are some sufficient statistics such that $\mathbf{t}(\mathbf{y}) = (t_1(\mathbf{y}), t_2(\mathbf{y}), \dots, t_k(\mathbf{y}))^T (\in \Theta)$. For simplicity of notation, $\mathbf{t}(\mathbf{y})$ and $\mathbf{s}(\mathbf{y})$ will be written in short as \mathbf{t} and \mathbf{s} , respectively. Note that $\pi(\boldsymbol{\theta} | \mathbf{y}, H_a)$ and $B(\mathbf{y})$, and thus the Bayes rule described in Theorem 1, are functions of \mathbf{y} only through (\mathbf{t}, \mathbf{s}) .

We further assume that ν is permutation invariant, and that the prior is such that the posterior $\pi(\boldsymbol{\theta} | \mathbf{y}, H_a)$ is decreasing in transposition (DT); see Hollander, Proschan and Sethuraman (1977). In other words, we assume that there exists $\mathbf{t}_* = (t_1^*, t_2^*, \dots, t_k^*)$, a function of (\mathbf{t}, \mathbf{s}) , such that $\pi(\boldsymbol{\theta} | \mathbf{y}, H_a)$, if denoted by $q(\boldsymbol{\theta}, \mathbf{t}_*; \mathbf{s})$, satisfies the following conditions

- (a) $q(\rho(\boldsymbol{\theta}), \rho(\mathbf{t}_*); \mathbf{s}) = q(\boldsymbol{\theta}, \mathbf{t}_*; \mathbf{s})$ for all permutations ρ .
- (b) $q(\boldsymbol{\theta}^{(i,j)}, \mathbf{t}_*; \mathbf{s}) \leq q(\boldsymbol{\theta}, \mathbf{t}_*; \mathbf{s})$, whenever $t_i^* \geq t_j^*$, and $\theta_i \geq \theta_j$, where $\boldsymbol{\theta}^{(i,j)}$ denotes the vector $\boldsymbol{\theta}$ with its i^{th} and j^{th} components interchanged.

We shall say in this case that the posterior is DT in $(\boldsymbol{\theta}, \mathbf{t}_*)$.

From Bahadur-Goodman-Lehmann-Eaton Theorem (Bahadur and Goodman, 1952; Lehmann, 1966; Eaton, 1967), it follows that $E(L(\boldsymbol{\Xi}, i) | \mathbf{y}, H_a) \leq E(L(\boldsymbol{\Xi}, j) | \mathbf{y}, H_a)$ if $t_i^* \geq t_j^*$ for $i, j \in \{1, 2, \dots, k\}$, $i \neq j$. This implies that $\min_{i=1,2,\dots,k} E[L(\boldsymbol{\Xi}, i) | \mathbf{y}, H_a]$ is attained at a component that corresponds to the $\max_{j=1,2,\dots,k} t_j^*$. Thus, from Theorem 1, we get the following result.

Theorem 2 *Let $[k]_{\mathbf{t}_*}$ denotes the index at which t_j^* ($j = 1, 2, \dots, k$) is maximum. If the posterior $\pi(\boldsymbol{\theta} | \mathbf{y}, H_a)$ is DT in $(\mathbf{t}_*, \boldsymbol{\theta})$, then under the Assumption A, the Bayes rule under the prior (3) is given by*

Accept the null hypothesis $H_0 : \theta_1 = \theta_2 = \dots = \theta_k$ if

$$E[(L(\boldsymbol{\Xi}, 0) - L(\boldsymbol{\Xi}, [k]_{\mathbf{t}_*})) | \mathbf{y}, H_a] \leq \frac{p}{1-p} B(\mathbf{y})(l_1 - l_0). \quad (13)$$

Otherwise select $H_j : \theta_j = \theta_{[k]} > \theta_{[1]}$ that corresponds to the index $[k]_{\mathbf{t}_}$.*

If ties occur then select one of the associated H_j with equal probability.

We now discuss the probability models and the priors that leads to the DT property of the posterior $\pi(\boldsymbol{\theta} | \mathbf{y}, H_a)$. It is easy to see that if the prior $g_a(\boldsymbol{\theta} | \alpha)$ is non-informative, i.e., it is permutation symmetric in $\boldsymbol{\theta}$, then the posterior is DT in $(\mathbf{t}, \boldsymbol{\theta})$ if $h(\mathbf{t}, \mathbf{s}; \boldsymbol{\theta}, \alpha)$ is DT in $(\mathbf{t}, \boldsymbol{\theta})$, i.e., $h(\rho(\mathbf{t}), \mathbf{s}; \rho(\boldsymbol{\theta}), \alpha) = h(\mathbf{t}, \mathbf{s}; \boldsymbol{\theta}, \alpha)$ for all permutations $\rho(\cdot)$, and $h(\mathbf{t}^{(i,j)}, \mathbf{s}; \boldsymbol{\theta}, \alpha) \leq h(\mathbf{t}, \mathbf{s}; \boldsymbol{\theta}, \alpha)$ whenever $t_i \geq t_j$, and $\theta_i \geq \theta_j$ for all α and \mathbf{s} . In this case Theorem 2 hold for $\mathbf{t}_* = \mathbf{t}$.

Many densities have the DT property; for example, densities for the variance balanced design under a normal and elliptical linear model (Bansal and Gupta, 1997;

Bansal, Misra and van der Meulen, 1997), and densities for a balanced design under generalized linear model (Bansal and Miescke, 2006), and densities with property M (Eaton, 1967).

Consider the density of the form

$$f(\mathbf{y}; \boldsymbol{\theta}, \alpha) = \exp\{(\mathbf{t}^T \boldsymbol{\theta} - b(\boldsymbol{\theta}))/\sigma(\alpha) - c(\mathbf{s}, \alpha)\}r(\mathbf{y}) \quad (14)$$

where, $b(\boldsymbol{\theta})$ is a permutation symmetric function of $\boldsymbol{\theta}$, and $\sigma(\alpha)$ and $c(\mathbf{s}, \alpha)$ are some functions. And consider the following conjugate prior density (Bansal and Miescke, 2006)

$$g_\gamma(\boldsymbol{\theta}|\alpha) = \exp\{(\boldsymbol{\gamma}^T \boldsymbol{\theta} - \kappa(\boldsymbol{\gamma}))/\sigma(\alpha)\}g_*^{(1)}(\boldsymbol{\theta}|\alpha) \quad (15)$$

where given α , $g_*^{(1)}(\boldsymbol{\theta}|\alpha)$ is a permutation symmetric ν -density of $\boldsymbol{\theta}$, $\boldsymbol{\gamma} \in R^k$ is a vector of hyperparameters, and $\kappa(\boldsymbol{\gamma})$ is a permutation symmetric function in $\boldsymbol{\gamma}$. Note that the prior information on the components of $\boldsymbol{\theta}$ can be conveyed through the hyperparameters $(\gamma_1, \gamma_2, \dots, \gamma_k)$, see Bansal and Miescke (2006).

Lemma 3 *For the density (14), under the conjugate prior (15), the posterior $\pi(\boldsymbol{\theta}|\mathbf{y}, H_\alpha)$ is DT in $(\mathbf{t}_*, \boldsymbol{\theta})$, where $\mathbf{t}_* = (t_1 + \gamma_1, t_2 + \gamma_2, \dots, t_k + \gamma_k)^T$.*

Proof. *From (14) and (15), note that the posterior density*

$$\pi(\boldsymbol{\theta}|\mathbf{y}, H_\alpha) \propto \int \exp\{((\mathbf{t} + \boldsymbol{\gamma})^T \boldsymbol{\theta} - b(\boldsymbol{\theta}) - \kappa(\boldsymbol{\gamma}))/\sigma(\alpha) - c(\mathbf{s}, \alpha)\}g_*^{(1)}(\boldsymbol{\theta}|\alpha)\omega(\alpha)d\nu_1(\alpha).$$

Proof now follows from the fact that $b(\boldsymbol{\theta})$ and $g_^{(1)}(\boldsymbol{\theta}|\alpha)$ are permutation symmetric in $\boldsymbol{\theta}$. ■*

The assumption above that $b(\boldsymbol{\theta})$ is permutation symmetric amounts to the assumption that the design is balanced. However, posteriors under unbalanced design can also have DT property as we discuss it for the one parameter exponential families in Section 4; see also Abughalous and Bansal (1995).

3 Bayes Rules Under Specific Loss Functions

In this section, we discuss the Bayes rules under two different loss functions, the "0-1" loss and a linear loss, as defined below. The loss functions presented here are only few examples; different variants of these loss functions can be considered, for example,

by replacing $\bar{\theta}$ in L_2 below by $\theta_{[1]} = \min(\theta_1, \theta_2, \dots, \theta_k)$. In practice, the choice of the loss function may depend on the problem of interest or the experimenter's preference. The theorems presented here are special cases of Theorem 1 and Theorem 2, and thus are given without proofs.

3.1 "0-1" loss function L_1

$$L_1(\boldsymbol{\theta}, 0) = \begin{cases} 0 & \text{if } \boldsymbol{\theta} \in \Theta_0 \\ 1 & \text{if } \boldsymbol{\theta} \notin \Theta_0 \end{cases}$$

and for $i = 1, 2, \dots, k$,

$$L_1(\boldsymbol{\theta}, i) = \begin{cases} 0 & \text{if } \boldsymbol{\theta} \notin \Theta_0 \text{ and } \theta_i = \theta_{[k]} \\ 1, & \text{otherwise} \end{cases}$$

Under this loss, clearly $l_0 = 0$, and $l_1 = 1$, and it is easy to see that

$$E[(L(\boldsymbol{\Xi}, 0) - L(\boldsymbol{\Xi}, i)) | \mathbf{y}, H_a] = P(\Xi_i = \Xi_{[k]} | \mathbf{y}, H_a) = p_i(\mathbf{y}), \text{ say.}$$

Theorem 4 *Under the "0-1" loss L_1 , the Bayes rule is given as follows.*

Accept H_0 if for all $i = 1, \dots, k$,

$$p_i(\mathbf{y}) = P(\Xi_i = \Xi_{[k]} | \mathbf{y}, H_a) \leq \frac{p}{1-p} B(\mathbf{y}). \quad (16)$$

Otherwise, select $H_j : \theta_j = \theta_{[k]} > \theta_{[1]}$ if

$$p_j(\mathbf{y}) = \max_{i=1,2,\dots,k} p_i(\mathbf{y}). \quad (17)$$

If, in addition, the posterior $\pi(\boldsymbol{\theta} | \mathbf{y}, H_a)$ is DT in $(\boldsymbol{\theta}, \mathbf{t}_)$, then, provided H_0 is rejected, the selection of H_j can be made according to the largest of t_i^* , $i = 1, 2, \dots, k$.*

We also recommend in the spirit of the Bayesian hypothesis methodology that one should report $p_0(\mathbf{y})$ when the null hypothesis H_0 is accepted and $p_j(\mathbf{y})$ when H_j is selected, c.f., Berger (2003) and Berger, Brown and Wolpert (1994). Note that $p_0(\mathbf{y})$ is the posterior probability of H_0 , and $p_j(\mathbf{y})$ is the probability that a posteriori $\theta_j = \theta_{[k]}$. It may also be useful to report the quantity $r_{L_1} = p_j(\mathbf{y}) / \max_{i \neq j} p_i(\mathbf{y})$ if H_j is selected, which would compare the selected component with rest of the components. The

significant high value of this would give an evidence regarding the strength of the selected component as compared to the other components.

Remark 2 *The above theorem implies that if the Bayes factor $B(\mathbf{y}) > 1$, and if the prior odds of accepting H_0 , i.e., $p/(1-p) = 1$, then H_0 will be accepted since $0 \leq p_i(\mathbf{y}) \leq 1$ for all $i = 1, 2, \dots, k$. This is in line with the Bayesian hypothesis testing. On the other hand, if $B(\mathbf{y}) < 1/k$ and the prior odds of accepting H_0 is 1, then (16) cannot hold for all $i = 1, 2, \dots, k$ since $\sum_{i=1}^k p_i(\mathbf{y}) = 1$, and thus, necessarily, one of the H_j , $j = 1, 2, \dots, k$, that corresponds to the largest of $p_i(\mathbf{y})$, $i = 1, 2, \dots, k$, will be selected. Also note that if a posteriori most but few of the θ_i s are away from $\theta_{[k]}$ in the sense that $p_i(\mathbf{y})$ s are very low, then $\max_{i=1, \dots, k} p_i(\mathbf{y})$ will be high since $\sum_{i=1}^k p_i(\mathbf{y}) = 1$, and thus the rejection of H_0 would be highly likely.*

3.2 Linear Loss L_2

$$L_2(\boldsymbol{\theta}, 0) = \begin{cases} l_0 & \text{if } \boldsymbol{\theta} \in \Theta_0 \\ \theta_{[k]} - \bar{\theta} & \text{if } \boldsymbol{\theta} \notin \Theta_0 \end{cases}$$

and for $i = 1, 2, \dots, k$,

$$L_2(\boldsymbol{\theta}, i) = \begin{cases} l_1 & \text{if } \boldsymbol{\theta} \in \Theta_0 \\ \theta_{[k]} - \theta_i & \text{if } \boldsymbol{\theta} \notin \Theta_0 \end{cases}$$

where l_0 and l_1 are some constants which should be appropriately defined according to the scale of θ 's.

Theorem 5 *Under the linear loss L_2 , the Bayes rule is given as follows.*

Accept H_0 if for all $i = 1, \dots, k$,

$$E[(\Xi_i - \bar{\Xi}) | \mathbf{y}, H_a] \leq \frac{p}{1-p} B(\mathbf{y})(l_1 - l_0). \quad (18)$$

Otherwise, select $H_j : \theta_j = \theta_{[k]} > \theta_{[1]}$ according to the largest of

$$E[\Xi_i | y, H_a], \quad i = 1, 2, \dots, k.$$

If the posterior $\pi(\boldsymbol{\theta} | \mathbf{y}, H_a)$ is DT in $(\boldsymbol{\theta}, \mathbf{t}_)$, then, provided H_0 is rejected, the selection of H_j can be made according to the largest of t_i^* , $i = 1, 2, \dots, k$.*

As we have recommended for the "0-1" loss function, for the purpose of reporting the strength of evidence in favor of accepting H_0 or in favor of selecting H_j , here we recommend reporting $p_0(\mathbf{y})$ when H_0 is accepted, and reporting $s_{L_2}(\mathbf{y}) = (E[\Xi_j | \mathbf{y}] - \min_{i \neq j} E[\Xi_i | \mathbf{y}]) / \zeta(\mathbf{y})$ when H_j is selected, where $\zeta(\mathbf{y})$ is some normalizing constants, for example $\zeta(\mathbf{y}) = (\sum_{i=1}^k E[(\Xi_i - \bar{\Xi})^2 | \mathbf{y}])^{\frac{1}{2}}$. A high value of this would give an evidence about how further away the selected component is from rest of the components.

4 One-Parameter Exponential Families of Distributions

In this section, we consider k populations with respective densities (with respect to a σ -finite measure ζ)

$$f(y|\theta_i) = \exp\{y\theta_i - M(\theta_i)\}h(y), \quad i = 1, \dots, k \quad (19)$$

where $h(y)$ is a density with the cumulant generating function $M(\cdot)$, θ_i is a natural parameter belonging to the convex set $\{\theta : M(\theta) < \infty\}$. Let $\mu_i = E[X_i|\theta_i]$ be the mean of the i^{th} population. The problem of interest is to test $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ against selecting one of k alternatives $H_i : \mu_i = \mu_{[k]} > \mu_{[1]}$, $i = 1, 2, \dots, k$. We shall assume that $\mu_i = M'(\theta_i)$ is a one-to-one increasing function of θ_i . Thus the problem is equivalent to testing $H_0 : \theta_1 = \theta_2 = \dots = \theta_k$ against selecting one of k alternatives $H_i : \theta_i = \theta_{[k]} > \theta_{[1]}$, $i = 1, 2, \dots, k$.

Since, in practice, hypothesis problems would be stated in terms of μ_i , $i = 1, 2, \dots, k$, it would be important to define prior in terms of μ_i , $i = 1, 2, \dots, k$. However, since the conjugate priors for the exponential families are easy to express in terms of θ_i , $i = 1, 2, \dots, k$, we would define the prior in terms of $\theta_1, \theta_2, \dots, \theta_k$. The prior for $\mu_1, \mu_2, \dots, \mu_k$ then can be interpreted from the transformation $\mu_i = M'(\theta_i)$. The relative prior information about the components $\mu_1, \mu_2, \dots, \mu_k$ can be obtained from the relative prior information on $\theta_1, \theta_2, \dots, \theta_k$. For example, the relative prior information conveyed on θ_1 and θ_2 via $\lambda_1 > \lambda_2$, where λ_i s are the hyperparameters as defined below, will be same for the components μ_1 and μ_2 as well.

For the prior $g_a(\boldsymbol{\theta})$ under H_a , we assume that $\theta_1, \theta_2, \dots, \theta_k$ are apriori independent, and θ_i follows the conjugate prior $\zeta(\varphi_i, \lambda_i)$ (Diaconis and Yalvisakar, 1979) with density

$$g(\theta_i) = \kappa(\varphi_i, \lambda_i) \exp\{\varphi_i(\lambda_i\theta_i - M(\theta_i))\}, \quad i = 1, 2, \dots, k, \quad (20)$$

where λ_i and $\varphi_i > 0$, $i = 1, 2, \dots, k$, are the hyperparameters, and $\kappa(\varphi_i, \lambda_i)$, $i = 1, 2, \dots, k$, are the normalizing constants. The values of (φ_i, λ_i) , $i = 1, 2, \dots, k$, can be chosen according to the prior information available or can be estimated empirically. Similarly, for the prior $g_0(\boldsymbol{\theta})$ under H_0 , we assume that the common value $\theta_1 = \theta_2 = \dots = \theta_k = \theta$ (say) follows the conjugate prior $\zeta(\varphi_0, \lambda_0)$ with density

$$g_0(\theta) = \kappa(\varphi_0, \lambda_0) \exp\{\varphi_0(\lambda_0\theta - M(\theta))\} \quad (21)$$

where $\varphi_0 > 0$ and λ_0 are the hyperparameters associated with the null hypothesis.

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})$ be a random sample of size n_i from the i^{th} population. Suitable sample sizes n_i , $i = 1, 2, \dots, k$, can be chosen so that the posterior is DT as we see below. It is easy to see that a posteriori under H_a , $\theta_1, \theta_2, \dots, \theta_k$ are independent, and $\theta_i | \mathbf{y}, H_a \sim \zeta(\varphi_i + n_i, \bar{y}_i^{(\varphi_i, \lambda_i)})$, $i = 1, 2, \dots, k$, where $\bar{y}_i^{(\varphi_i, \lambda_i)} = (n_i \bar{y}_i + \varphi_i \lambda_i) / (\varphi_i + n_i)$ is the posterior mean of θ_i . Similarly, under H_0 , the posterior distribution of the common θ is $\zeta(N + \varphi_0, \bar{y}_0^{(\varphi_0, \lambda_0)})$, where $N = \sum n_i$, and $\bar{y}_0^{(\varphi_0, \lambda_0)} = (\sum n_i \bar{y}_i + \lambda_0 \varphi_0) / (\varphi_0 + N)$. It can also be seen by computing the marginal densities $m_0(\mathbf{y})$ and $m_a(\mathbf{y})$ under H_0 and H_a , respectively that the Bayes factor is given by

$$B(\mathbf{y}) = \frac{m_0(\mathbf{y})}{m_a(\mathbf{y})} = \frac{\kappa(\varphi_0, \lambda_0) \prod_{i=1}^k \kappa(\varphi_i + n_i, \bar{y}_i^{(\varphi_i, \lambda_i)})}{\left(\prod_{i=1}^k \kappa(\varphi_i, \lambda_i) \right) \kappa(\varphi_0 + N, \bar{y}_0^{(\varphi_0, \lambda_0)})} \quad (22)$$

Now, from Theorem 1, we get the following result.

Theorem 6 *For k independent exponential families of distributions (19), when the priors are given by (20) and (21) under H_a and H_0 , respectively, then the Bayes rule accepts H_0 , if for all $i = 1, 2, \dots, k$,*

$$E[(L(\boldsymbol{\mu}, 0) - L(\boldsymbol{\mu}, i)) | \mathbf{y}, H_a] \leq \frac{p}{1-p} B(\mathbf{y})(l_0 - l_1), \quad (23)$$

where $B(\mathbf{y})$ is given by (22). Otherwise select H_j according to the largest of $E[L(\boldsymbol{\mu}, i) | \mathbf{y}, H_a]$, $i = 1, 2, \dots, k$, where expectations are with respect to the posterior distributions of θ_i , $i = 1, 2, \dots, k$ which are independent $\zeta(\varphi_i + n_i, \bar{y}_i^{(\varphi_i, \lambda_i)})$, $i = 1, 2, \dots, k$.

If the sample sizes n_i , $i = 1, 2, \dots, k$, are chosen such that $\varphi_i + n_i = n_\varphi$ (a constant for all i), $i = 1, 2, \dots, k$, then, from Lemma 3, the posterior is DT in $(\boldsymbol{\theta}, \bar{\mathbf{y}}^{(\varphi, \lambda)})$,

where $\bar{\mathbf{y}}^{(\varphi, \lambda)} = (\bar{y}_1^{(\varphi_1, \lambda_1)}, \bar{y}_2^{(\varphi_2, \lambda_2)}, \dots, \bar{y}_k^{(\varphi_k, \lambda_k)})^T$. In this case, from Theorem 2, we get the following result.

Corollary 6.1 *If n_i , $i = 1, 2, \dots, k$, are chosen such that $\varphi_i + n_i = n_\varphi$ for all $i = 1, 2, \dots, k$, where n_φ is some constant, then the Bayes rule accepts H_0 if (23) is satisfied for all i ; otherwise selects H_j that corresponds to the largest of $\bar{y}_m^{(\varphi_1, \lambda_1)}$, $m = 1, 2, \dots, k$.*

The Bayes rule is easy under the linear loss L_2 : $L_2(\boldsymbol{\mu}, 0) = \mu_{[k]} - \bar{\mu}$, and for $i = 1, 2, \dots, k$, $L_2(\boldsymbol{\mu}, i) = \mu_{[k]} - \mu_i$. Since $E[\mu_i | \mathbf{y}, H_a] = \bar{y}_i^{(\varphi_i, \lambda_i)}$ (Diaconis and Yalvisaker, 1979), from Theorem 6, the Bayes rule, under the loss L_2 , accepts H_0 , if for all $i = 1, 2, \dots, k$, $\bar{y}_i^{(\varphi_i, \lambda_i)} - \bar{y}^{(\varphi, \lambda)} \leq p/(1-p)B(\mathbf{y})(l_0 - l_1)$, where $B(\mathbf{y})$ is given by (22), and $\bar{y}^{(\varphi, \lambda)}$ is the average of $\bar{y}_i^{(\varphi_i, \lambda_i)}$, $i = 1, 2, \dots, k$. Otherwise it selects H_j according to the largest of $\bar{y}_i^{(\varphi_i, \lambda_i)}$, $i = 1, 2, \dots, k$.

Since, the computation of the posterior probabilities $P[\mu_i = \mu_{[k]} | \mathbf{y}, H_a]$, $i = 1, 2, \dots, k$, is not straight forward, a closed form solution of the Bayes rule is not readily available for the "0-1" loss function L_1 . However, the posterior probabilities can be computed in a single integral as

$$P[\mu_i = \mu_{[k]} | \mathbf{y}, H_a] = \int \prod_{j \neq i} F_{\mu_j | \mathbf{y}}(t) f_{\mu_i | \mathbf{y}}(t) dt,$$

where $F_{\mu_j | \mathbf{y}}(\cdot)$ and $f_{\mu_j | \mathbf{y}}(\cdot)$ denote the distribution function and the density function respectively of μ_j given \mathbf{y}, H_a .

5 Normal Populations

In this section, we discuss the Bayes rules under the "0-1" loss L_1 and the linear loss L_2 for independent normal populations $N(\theta_i, \sigma^2)$, $i = 1, 2, \dots, k$, when σ^2 is known or unknown. The known σ^2 case is a special case of Section 3. However, since we use the non-informative improper prior instead of the conjugate prior, we present this case separately here. We will present the known σ^2 and the unknown σ^2 cases simultaneously; however, the frequentists' simulation comparisons of the Bayes rules with the classical procedures will be made only for the known σ^2 . The main point of the simulation is to show that the power of the Bayes decision rules is better than the power of the classical rules.

The objective is to test $H_0 : \theta_1 = \theta_2 = \dots = \theta_k$ against selecting one of $H_i : \theta_i = \theta_{[k]} > \theta_{[1]}$, $i = 1, 2, \dots, k$. Suppose, n_i independent observations y_{ij} , $j = 1, 2, \dots, n_i$ are available from $N(\theta_i, \sigma^2)$, $i = 1, 2, \dots, k$, populations. In order to derive the Bayes rules of Section 3, we need the posterior distribution under H_0 and under H_a .

Consider the following non-informative improper priors

$$g_0(\theta)d\theta = \begin{cases} d\theta & \text{when } \sigma^2 \text{ is known} \\ \frac{1}{\sigma^2} d\theta d\sigma^2 & \text{when } \sigma^2 \text{ is unknown} \end{cases} \quad \text{under } H_0 \quad (24)$$

and

$$g_a(\boldsymbol{\theta})d\boldsymbol{\theta} = \begin{cases} d\boldsymbol{\theta} & \text{when } \sigma^2 \text{ is known} \\ \frac{1}{\sigma^2} d\boldsymbol{\theta} d\sigma^2 & \text{when } \sigma^2 \text{ is unknown} \end{cases} \quad \text{under } H_a \quad (25)$$

It is easy to see that when σ^2 is known, aposteriori $\theta | \mathbf{y} \sim N(\bar{y}, \sigma^2/N)$ under H_0 , and $\theta_i | \mathbf{y} \sim N(\bar{y}_i, \sigma^2/n_i)$, $i = 1, 2, \dots, k$, are independent under H_a ; and when σ^2 is unknown, aposteriori $\theta | \mathbf{y}, \sigma^2 \sim N(\bar{y}, \sigma^2/N)$ and $(\sigma^2)^{-1} | \mathbf{y} \sim (T)^{-1} \chi_{N-1}^2$ under H_0 , and $\theta_i | \mathbf{y}, \sigma^2 \sim N(\bar{y}_i, \sigma^2/n_i)$, $i = 1, 2, \dots, k$, are independent, and $(\sigma^2)^{-1} | \mathbf{y} \sim (W)^{-1} \chi_{N-k}^2$ under H_a . Here $N = \sum n_i$, $\bar{y}_i = \sum y_{ij}/n_i$, $\bar{y} = N^{-1} \sum \sum y_{ij}$, $T = \sum \sum (y_{ij} - \bar{y})^2$, and $W = \sum \sum (y_{ij} - \bar{y}_i)^2$. The Bayes factor $B(\mathbf{y}) = m_0(\mathbf{y})/m_1(\mathbf{y})$ is given by

$$B(\mathbf{y}) = \frac{\prod_{i=1}^k n_i^{1/2}}{(2\pi\sigma^2)^{(k-1)/2} N^{1/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2\right\}, \quad (26)$$

when σ^2 is known, and

$$B(\mathbf{y}) = \frac{\Gamma(\frac{N-1}{2}) \prod_{i=1}^k n_i^{1/2}}{\pi^{(k-1)/2} \Gamma(\frac{N-k}{2}) N^{1/2}} \frac{W^{-(k-1)/2}}{\left[1 + \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 / W\right]^{(N+1)/2}}, \quad (27)$$

when σ^2 is unknown.

Remark 3 *It is easy to see that for the priors (25), the posterior is DT in $(\boldsymbol{\theta}, \bar{\mathbf{y}})$ if and only if $n_1 = n_2 = \dots = n_k$, where $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k)^T$. In that case, the Bayes rule, from Theorem 2 under any loss that satisfies Assumption A, will select H_j according*

to the largest of \bar{y}_i , $i = 1, 2, \dots, k$, provided H_0 is rejected.

We now consider the Bayes rules under the "0-1" loss L_1 and the linear loss L_2 . For simplicity, we only consider the known σ^2 case. The unknown σ^2 case can be solved from the known σ^2 case by further taking the expectation of the posterior expected loss with respect to the posterior distribution of σ^2 which is given by $(\sigma^2)^{-1} | \mathbf{y} \sim (W)^{-1} \chi_{N-k}^2$.

We first consider the "0-1" loss L_1 . Since the posterior distribution of θ_i is $N(\bar{y}_i, \sigma^2/n_i)$, $i = 1, 2, \dots, k$, and since a posteriori they are independent,

$$\begin{aligned} p_i(\mathbf{y}) &= P(\Xi_i = \Xi_{[k]} | \mathbf{y}, H_a) \\ &= P(\Xi_j \leq \Xi_i \text{ for all } j \neq i | \mathbf{y}, H_a) \\ &= \int_{-\infty}^{\infty} \prod_{j \neq i} \Phi(\sqrt{n_j}(\frac{t - \bar{y}_j}{\sigma})) \frac{\sqrt{n_i}}{\sigma} \phi(\sqrt{n_i}(\frac{t - \bar{y}_i}{\sigma})) dt, \end{aligned} \quad (28)$$

where Φ and ϕ are the cumulative distribution function and the density function of the standard normal distribution, respectively.

Thus, from Theorem 4, under the "0-1" loss function L_1 , the Bayes rule accepts H_0 if for all $i = 1, 2, \dots, k$,

$$p_i(\mathbf{y}) \leq \frac{p}{1-p} B(\mathbf{y}). \quad (29)$$

Otherwise it selects $H_j : \theta_j = \theta_{[k]} > \theta_{[1]}$ according to the largest of $p_i(\mathbf{y})$, $i = 1, 2, \dots, k$, where $p_i(\mathbf{y})$, $i = 1, 2, \dots, k$, are given by (28), and $B(\mathbf{y})$ is given by (26).

For the Bayes rule under the loss L_2 , note that $E(\Xi_i - \bar{\Xi} | \mathbf{y}, H_a) = \bar{y}_i - \bar{y}$. Thus, from Theorem 5, the Bayes rule under the linear loss L_2 accepts H_0 , if for all $i = 1, 2, \dots, k$,

$$\bar{y}_i - \bar{y} \leq \frac{p}{1-p} B(\mathbf{y})(l_1 - l_0). \quad (30)$$

Otherwise it selects $H_j : \theta_j = \theta_{[k]} > \theta_{[1]}$ according to the largest of \bar{y}_i , $i = 1, 2, \dots, k$.

5.1 Frequentists' Comparison

The Bayes rules given above are applicable only in Bayesian settings when the prior information about the null probability p is available. However, a frequentist's version of the above rules can be obtained by constraining them to be size γ tests ($0 < \gamma < 1$) in the following way. We assume for simplicity $n_1 = n_2 = \dots = n_k = n$ (say).

Accept H_0 if for all $i = 1, \dots, k$,

$$p_i(\mathbf{y}) \leq c_\gamma(k, n, \sigma)B(\mathbf{y}). \quad (31)$$

Otherwise, select $H_j : \theta_j = \theta_{[k]} > \theta_{[1]}$ according to the largest of \bar{y}_i , $i = 1, 2, \dots, k$, where $c_\gamma(k, n, \sigma)$ is a constant such that

$$\sup_{H_0} P(\max_{i=1, \dots, k} p_i(\mathbf{Y})/B(\mathbf{Y}) > c_\gamma(k, n, \sigma)) = \gamma. \quad (32)$$

Similarly, the size γ Bayes rule based on (30) can be defined as: Accept H_0 if for all $i = 1, \dots, k$,

$$\bar{y}_i - \bar{y} \leq d_\gamma(k, n, \sigma)B(\mathbf{y}). \quad (33)$$

Otherwise select $H_j : \theta_j = \theta_{[k]} > \theta_{[1]}$ according to the largest of \bar{y}_i , $i = 1, 2, \dots, k$, where $d_\gamma(k, n, \sigma)$ is a constant such that

$$\sup_{H_0} P(\max_{i=1, \dots, k} (\bar{Y}_i - \bar{Y})/B(\mathbf{Y}) > d_\gamma(k, n, \sigma)) = \gamma. \quad (34)$$

In order to find constants $c_\gamma(k, n, \sigma)$ and $d_\gamma(k, n, \sigma)$, note that, from (26) and (28), under $H_0 : \theta_1 = \theta_2 = \dots = \theta_k = \theta$, $P(\max_{i=1, \dots, k} p_i(\mathbf{Y})/B(\mathbf{Y}) > c_\gamma(k, n, \sigma))$ and $P(\max_{i=1, \dots, k} (\bar{Y}_i - \bar{Y})/B(\mathbf{Y}) > d_\gamma(k, n, \sigma))$ are independent of θ . Thus, without loss of generality, we can assume that $\theta = 0$. Furthermore, it can be also seen that $c_\gamma(k, n, \sigma) = (\sigma^2/n)^{(k-1)/2}c_\gamma(k, 1, 1)$, and $d_\gamma(k, n, \sigma) = (\sigma^2/n)^{k/2}d_\gamma(k, 1, 1)$. Here, $c_\gamma(k, 1, 1)$ and $d_\gamma(k, 1, 1)$ are the upper γ -cutoff points of the distributions of $\max_{i=1, \dots, k} p_i(\mathbf{Z})/B(\mathbf{Z})$ and $\max_{i=1, \dots, k} (Z_i - \bar{Z})/B(\mathbf{Z})$, respectively, where $\mathbf{Z} = (Z_1, Z_2, \dots, Z_k)^T$ and Z_1, Z_2, \dots, Z_k are the i.i.d $N(0, 1)$ random variables. The coefficients $c_\gamma(k, 1, 1)$ and $d_\gamma(k, 1, 1)$, thus, can be obtained by simulating the $N(0, 1)$ random variables.

In our simulations, we find the coefficients $c_\gamma(k, 1, 1)$ and $d_\gamma(k, 1, 1)$ by simulating 10,000 copies of Z_1, Z_2, \dots, Z_k , and then by taking the upper γ -cutoff points of the empirical distributions of $\max_{i=1, \dots, k} p_i(\mathbf{Z})/B(\mathbf{Z})$ and $\max_{i=1, \dots, k} (Z_i - \bar{Z})/B(\mathbf{Z})$, respectively. It should be noted that the computation of $c_\gamma(k, 1, 1)$ is time consuming due to the numerical computation of the integral in (28), especially for large k . All of the simulations and computations were done on *Matlab R2006a* (The MathWorks, 2006).

We compare the Bayes rules given above against the classical rule based on the likelihood ratio tests. When σ^2 is known, the likelihood ratio test rejects H_0 if $\chi^2 = \sum n_i(\bar{y}_i - \bar{y})^2/\sigma^2 > \chi_\gamma^2$. And when H_0 is rejected, the classical rule selects H_i according to the largest of \bar{y}_i , $i = 1, 2, \dots, k$. The results are presented in Tables 1-3. The results presented are for three cases of $k = 20, 40$ and 100 , and two sample sizes of $n = 5$ and $n = 25$. The tables present the powers of the rejection regions of the "0-1" loss Bayes rule, linear loss Bayes rule, and the classical rule based on $\gamma = 0.05$. The conditional powers of correctly selecting the best population conditioned upon rejecting the null hypothesis are also presented. All the power calculations were done based on simulation of 1000 samples. The computation of $p_i(y)$, $i = 1, \dots, k$, was very time consuming, especially for large k . For this reason, only the linear loss was considered for the case $k = 100$. Without loss of generality, the correct population with the largest mean was taken to be the first population. The configurations of $\theta_1, \theta_2, \dots, \theta_k$ are based on the partition in such a way that a unique population has the highest value of the θ_i s, while some (about 10% or 25%) have the second highest values of the θ_i s, and most of them (the remaining θ_i s) have the 0 values. For example, $(\theta_1, \theta_2, \dots, \theta_k)$ with a partition $(1.0^{\{1\}}, 0.3^{\{9\}}, 0.0^{\{90\}})$ means $\theta_1 = 1.0$, $\theta_2 = \dots = \theta_{10} = 0.3$, and $\theta_{11} = \dots = \theta_{100} = 0.0$. The reason for choosing such configurations was that in a high dimensional data, such as in microarray, only few values are expressed with significantly high values.

It was observed consistently that the power of the "0-1" loss Bayes rule was higher than that of the linear loss Bayes or the classical rule. Also the power of the linear loss Bayes rule was consistently higher than the power of the classical rule except in some rare cases. The power of the "0-1" loss Bayes was, in some cases, more than 5% higher than that of the classical rule. The improvement does seem to vary from a small sample size ($n = 5$) to a large sample size ($n = 25$). It was also observed that, for the larger number of populations $k = 40$ or $k = 100$ as compared to $k = 20$, the improvement was significantly better. For example, for $n = 25$, for the configuration $(0.8^{\{1\}}, 0.1^{\{k/10-1\}}, 0.0^{\{9k/10\}})$, the power improvement for $k = 20$ was 3.7% (from 0.6500 to 0.6930), while for $k = 40$, it was 5.5% (from 0.5110 to 0.5660). The power of rejecting the null hypothesis increased as the sample size increased from $n = 5$ to $n = 25$ in all cases. It was also observed that the "0-1" loss Bayes rule performed consistently better in selecting the correct population when the null hypothesis was rejected. This is an important point since it may be more desirable to have a high

power of detecting the correct population once the null hypothesis is rejected than having a high power of rejecting the null hypothesis itself. The tables also include one configuration in the reverse direction for each of $k = 20, 40, 100$ and $n = 5, 25$ in which a high percentage of the populations have the highest mean value while vary few have low values. It is interesting to see that, in this case, the power of rejecting the null hypothesis for the classical rule is slightly better. The power of selection was not computed in these cases because there is no single best population. Overall the results show that, when most of the populations have low mean values and few have high mean values, the power of rejecting the null hypothesis under the classical rule is poor as compared to the "0-1" loss Bayes or the linear loss Bayes rule, and the classical rule selects the wrong populations more often (after rejecting the null hypothesis) than does the "0-1" loss Bayes or the linear loss Bayes rule.

6 Conclusion

In this paper, we have used the decision theoretic Bayesian methodology to demonstrate how the problem of selecting the best component, if there is a difference in the components, can lead to improved decision rule over the classical frequentist's rule. Classical tests are uniformly best invariant that distribute power uniformly in all directions. However, when the objective is to select the best component, then there is no need for rotational invariant procedures. For such problems better procedures can be obtained that have good power in certain directions. Although the current work is based on the Bayesian principle, the decision rules so obtained also enjoy the frequentists' optimality property. The performance of the Bayesian rules was observed to be consistently better than that of the classical rule under a certain type of configuration of the alternatives that pertain to the large dimensional data where only a minority of the variables have significantly high values. The present approach uses simultaneous testing and selection. There can be different approaches based on first testing and then selection or first selection then testing. Such procedures may be desirable if one wants different controlled errors for testing and selection.

One of the main points of this paper was to show that a decision theoretic approach to certain problems with appropriate loss functions can be more powerful than procedures based on an ad hoc classical approach. In other words, for the problems such as multiple hypotheses problems, it is better to define them in a decision theoretic

framework with appropriate loss functions. The Bayes decision rules so obtained will probably have better power than the classical ad hoc procedures under appropriate alternatives. For the problems related to microarray data analysis, where the purpose is to find not necessarily the best components but several components with high yields if they exist, the current methodology can perhaps be extended with loss functions defined appropriately. Such problems in the context of the selection problems, when comparing with a control, have been considered by several authors; see, for example, Tong (1969), Gupta and Hsiao (1983), Huang, Panchapakesan and Tseng (1984), and Gupta and Li (2005).

The aspect of choosing an appropriate prior was not explored here. If apriori a certain configuration is known about the population parameters, then an appropriate prior accommodating such information could yield a better Bayes rule. Another aspect of the prior that was not explored here was that the hierarchical structure of the prior (3) assumes a first stage prior only on the null hypothesis H_0 , but not on each individual H_i , $i = 1, \dots, k$. The methodology presented here can be extended easily to incorporate distinct first stage priors $\pi(H_i)$, $i = 1, 2, \dots, k$, on each individual H_i , $i = 1, \dots, k$. This may be necessary if available prior information on the ranking of H_j needs to be implemented.

The methods presented here can be applied to the problems in general design models as well as in the regression models. As an example, consider a randomized block design where it is of interest to test the hypothesis of no treatment effect, and upon rejection, to select the best treatment; see Bansal and Gupta (1997) and Bansal and Miescke (2002) in the context of the selection problem. For an application in a regression model, see Fong (1992).

Acknowledgement 7 *The author is greatly indebted to the two referees who were instrumental in making the presentation of the manuscript more concise and clear. One of the referees comments led to major improvements with some new results, especially the results of Section 5.1.*

References

- Abughalous, M.M., and Bansal, N.K. (1995). "On selecting the best natural exponential families with quadratic variance function," *Statist. Probab. Lett.*, 25, 341-349.

- Bahadur, R.R. (1950). "On the problem in the theory of k populations," *Ann. Math. Statist.*, 20, 362-375.
- Bahadur, R.R., and Goodman, L.A. (1952). "Impartial decision rules and sufficient statistics," *Ann. Math. Statist.*, 23, 553-562.
- Bansal, N.K., and Gupta, S. (1997). "On the natural rule in general linear models," *Metrika* 46, 59-69.
- Bansal, N.K., and Miescke, K.J. (2002). "Simultaneous selection and estimation in general linear models," *J. Statist. Plan. Inference*, 104, 377-390.
- Bansal, N.K., and Miescke, K. (2006). "On selecting the best treatment in a generalized linear model," *J. Statist. Plan. Inference*, 136, 2070-2086.
- Bansal, N.K., Misra, N. and van der Meulen, E.C. (1997). "On the minimax decision rules in ranking problems," *Statist. Probab. Lett.*, 34, 179-186.
- Bechhofer, R.E. (1954). "A single-sample multiple decision procedure for ranking means of normal populations with known variances," *Ann. Math. Statist.*, 25, 16-39.
- Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. Roy. Statist. Soc.*, Ser. B, 57, 289-300.
- Benjamini, Y. and Yekutieli, D. (2005). "False discovery rate-adjusted multiple confidence intervals for selected parameters," *J. Amer. Statist. Assoc.*, 100, 71-85.
- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.
- Berger, J.O. (2003). "Could Fisher, Jeffreys and Neyman have agreed on testing (with discussion)," *Statist. Sci.*, 18, 1-32.
- Berger, J.O., Brown, L.D., and Wolpert, R.L. (1994). "A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing," *Ann. Statist.*, 22, 1787-1807.

- 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. Amer. Statist. Assoc.*, 83, 364-373.
- Diaconis, P., and Ylvisaker, S. (1979). "Conjugate priors for exponential families," *Ann. Statist.*, 7, 269-281.
- Eaton, M.L. (1967). "Some optimum properties of ranking procedures," *Ann. Statist.*, 38, 124-137.
- Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). "Empirical Bayes analysis of a microarray experiment," *J. Amer. Statist. Assoc.*, 96, 1151-1160.
- Fong, D.K.H. (1992). "Ranking and estimation of related means in the presence of a covariate- a Bayesian approach," *J. Amer. Statist. Assoc.*, 87, 1128-1136.
- Gupta, S.S. (1956). "On a decision rule for a problem in ranking means," Mimeo, Ser. 150, Institute of Statistics, University of North Carolina, Chapel Hill, NC.
- Gupta, S.S., and Hsiao, P. (1983). "Empirical Bayes rules for selecting good populations," *J. Statist. Plan. Inference*, 8, 87-101.
- Gupta, S.S. and Li, J. (2005). "On empirical Bayes procedures for selecting good populations in a positive exponential family," *J. Statist. Plan. and Inference*, 129, 3-18.
- Huang, D.Y., Panchapakesan, S., and Tseng, S.T. (1984). "Some locally optimal subset selection rules for comparison with a control," *J. Statist. Plan. Inference*, 9, 63-72.
- Ishwaran, H. and Rao, J.S. (2003). "Detecting differentially expressed genes in microarrays using Bayesian model selection," *J. Amer. Statist. Assoc.*, 98, 438-455.
- Hollander, M., Proschan, F. and Sethuraman, J. (1977). "Functions decreasing in transposition and their applications in ranking problems," *Ann. Statist.*, 5, 722-733.
- Karlin, S. and Truax, D. (1960). "Slippage problems," *Ann. Math. Statist.*, 31, 296-324.

Lehmann, E.L. (1966). “On a theorem of Bahadur and Goodman,” *Ann. Statist.*, 37, 1-6.

Tong, Y.L. (1969). “On partitioning a set of normal populations by their locations with respect to a control,” *Ann. Math. Statist.*, 40, 1300-1324.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). “Significant analysis of microarrays applied to the ionizing radiation response,” *Proc. Natl. Acad. Sci. USA*, 98, 5116-5121.

Westfall, P.H., and Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p -Value Adjustment*. New York: John Wiley.