*Sankhyā : The Indian Journal of Statistics*
2006, Volume 68, Part 3, pp. 436-460


# Bayesian Maximum *a posteriori* Multiple Testing Procedure

Felix Abramovich
*Tel Aviv University, Tel Aviv, Israel*
Claudia Angelini
*Consiglio Nazionale delle Ricerche, Napoli, Italy*

**Abstract**

We consider a Bayesian approach to multiple hypothesis testing. A hierarchical prior model is based on imposing a prior distribution $\pi(k)$ on the number of hypotheses arising from alternatives (false nulls). We then apply the maximum *a posteriori* (MAP) rule to find the most likely configuration of null and alternative hypotheses. The resulting MAP procedure and its closely related step-up and step-down versions compare ordered Bayes factors of individual hypotheses with a sequence of critical values depending on the prior. We discuss the relations between the proposed MAP procedure and the existing frequentist and Bayesian counterparts. A more detailed analysis is given for the normal data, where we show, in particular, that by choosing a specific $\pi(k)$, the MAP procedure can mimic several known familywise error (FWE) and false discovery rate (FDR) controlling procedures. The performance of MAP procedures is illustrated on a simulated example.

*AMS* (2000) *subject classification.* Primary 62F15, 62F03.
*Keywords and phrases.* Bayes factor, false discovery rate, familywise error, hierarchical prior, maximum *a posteriori* rule, multiple hypothesis testing, $p$-value.

## 1 Introduction

Consider the standard multiple hypothesis testing set-up. Suppose that we have $n$ independent vectors of observations $\boldsymbol{Y}_i,\ i = 1, ..., n$ of sizes $m_i$, where $\boldsymbol{Y}_i \sim f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i)$, and $\boldsymbol{\theta}_i$ is a $d_i$-dimensional parameter vector in $\Omega_i \subset \mathbb{R}^{d_i}$. Given this data, we want to simultaneously test $n$ nonnested hypotheses

$$H_{0i} : \boldsymbol{\theta}_i \in \Theta_i \quad \text{vs.} \quad H_{1i} : \boldsymbol{\theta}_i \in \bar{\Theta}_i, \tag{1.1}$$

where $\Theta_i \cap \bar{\Theta}_i = \emptyset$, $\Theta_i \cup \bar{\Theta}_i = \Omega_i$, $i = 1, \ldots, n$. Following a frequentist approach to test $H_{0i}$ against $H_{1i}$, one chooses some test statistic $T_i(\boldsymbol{Y}_i)$, and the inference is typically based on the corresponding $p$-value $p_i$. Since there are $n$ such individual tests, the resulting $p$-values should be adjusted for multiplicity. The multiplicity correction depends on the type of the error one wishes to control. The traditional concern for multiplicity effect has been about the familywise error (FWE) – the probability of even a single Type I error in a series of $n$ tests. The widely known Bonferroni procedure is only an example, and more powerful FWE controlling procedures are currently available for many multiple testing problems (see Hochberg and Tamhane, 1987; Hsu, 1996 for reviews). Such a severe criterion implies however substantially reduced power properties especially when $n$ is large. A less stringent alternative to the FWE is the false discovery rate (FDR) criterion of Benjamini and Hochberg (1995). Unlike FWE, FDR controls the expected proportion of Type I errors among hypotheses being rejected (false discoveries) rather than the probability of even a single Type I error. The resulting FWE or FDR controlling procedures are typically stepwise in nature where the ordered $p$-values $p_{(1)} \leq \ldots \leq p_{(n)}$ are in effect compared with a series of properly chosen critical values. Step-up procedures start with testing the least significant hypothesis with the largest $p$-value $p_{(n)}$ and continue with decreasing $p$-values until the first rejection of the null hypothesis. Step-down procedures start with $p_{(1)}$ and continue with increasing $p$-values until the first acceptance.

A Bayesian approach to hypothesis testing in general and to the multiplicity problem in particular are conceptually different. Sarkar and Chen (2004) give an overview of the current Bayesian perspective on multiple testing. Consider first the $i$-th test as a single test. One assumes a prior distribution on $\boldsymbol{\theta}_i$ of the form

$$\pi_i(\boldsymbol{\theta}_i) = \begin{cases} \pi_{0i} p_{0i}(\boldsymbol{\theta}_i), & \text{if } \boldsymbol{\theta}_i \in \Theta_i \\ \pi_{1i} p_{1i}(\boldsymbol{\theta}_i), & \text{if } \boldsymbol{\theta}_i \in \bar{\Theta}_i, \end{cases}$$

where $p_{0i}(\boldsymbol{\theta}_i)$ and $p_{1i}(\boldsymbol{\theta}_i)$ are densities on $\Theta_i$ and $\bar{\Theta}_i$ respectively, and $\pi_{0i} + \pi_{1i} = 1$. The inference is then based on the posterior distribution $\pi_i(\boldsymbol{\theta}_i|\boldsymbol{Y}_i)$ according to the chosen loss. For the standard "0-1" loss, the null hypothesis $H_{0i}$ is rejected if the resulting posterior odds ratio $\pi_i(\boldsymbol{\theta}_i \in \Theta_i|\boldsymbol{Y}_i)/\pi_i(\boldsymbol{\theta}_i \in \bar{\Theta}_i|\boldsymbol{Y}_i)$ is smaller than 1 or, equivalently, if the corresponding Bayes factor $B_i$ is smaller than $\pi_{1i}/\pi_{0i}$. This Bayesian rule can be easily extended to a more general "0-$L_k$" loss under which the null hypothesis is rejected if $B_i < (L_{0i}/L_{1i})(\pi_{1i}/\pi_{0i})$. Turning back to multiple hypothesis testing, note

that from the Bayesian perspective, placing independent priors on $\boldsymbol{\theta}_i$ and using additive losses (e.g., a sum of "0-1" individual losses for each test) does not imply any multiplicity correction: the evident optimal Bayes rule in this case simply applies the corresponding individual Bayes testing rule to each one of $n$ tests "ignoring", therefore, the multiplicity effect. To account for multiplicity adjustments within a Bayesian framework, one should consider $\boldsymbol{\theta}_i$ to be *a priori* dependent to cause the posterior distribution of $\boldsymbol{\theta}_i$ to depend on all $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$. One possible way to introduce dependency among $\boldsymbol{\theta}_i$ is via *hierarchical* prior models. In situations where all likelihoods $f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i)$ have the same parametric form, like in multiple comparisons, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ are usually considered as an independent sample from a population distribution $\mathcal{F}(\boldsymbol{\theta})$ that may be parametric with possibly unknown parameters or even nonparametric (e.g., Waller and Duncan, 1969; Gopalan and Berry, 1998; Berry and Hochberg, 1999; Sarkar and Chen, 2004; Scott and Berger, 2006).

Generally, however, the likelihoods $f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i)$ may be of different forms, $\boldsymbol{\theta}_i$ may be of different dimensionalities, etc. and such type of hierarchical priors might be inappropriate. In this paper we propose a simple hierarchical prior model for a multiple hypothesis testing set-up (1.1) by imposing a prior distribution on the *number* of hypotheses arising from the alternatives (false nulls). We then apply the maximum *a posteriori* (MAP) rule and find the most plausible configuration of nulls and alternatives with the maximal posterior probability. Such a Bayes rule essentially corresponds to the following (non-additive) "nothing or everything" multiple "0-1" loss : the loss is zero if all $n$ hypotheses are inferred upon correctly and one if there is at least one wrong decision of any type regardless of their actual number. In the case of independent likelihoods, the resulting MAP procedure is based on the sequence of ordered individual Bayes factors somewhat similar in spirit to frequentist procedures operating with ordered $p$-values. We also present closely related step-up and step-down versions of the MAP multiple testing procedure and establish interesting parallels between them and several known frequentist and Bayesian multiple testing procedures.

The paper is organized as follows. The main results are presented in Section 2 where we propose the Bayesian MAP testing procedure and its stepwise versions. The relations between them and several existing Bayesian and frequentist multiple testing procedures are discussed in Section 3 while a more detailed study for the normal data is given in Section 4. In Section 5, we illustrate the performance of MAP procedures on a simulated example. Some concluding remarks are made in Section 6.

## 2 Main Results

*2.1. Hierarchical prior model.* Consider again the data set of $n$ independent $m_i$-dimensional vectors $\boldsymbol{Y}_i$, where $\boldsymbol{Y}_i \sim f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i),\ \boldsymbol{\theta}_i \in \Omega_i \subset \mathbb{R}^{d_i}$ and the multiple hypothesis testing problem (1.1). Following a Bayesian methodology, one needs to set prior odds simultaneously for each of $n$ null hypotheses. However, it might not be obvious to formulate prior beliefs on individual odds especially when they are not assumed to be independent. On the other hand, one has usually some intuition on the number (proportion) of hypotheses coming from the nulls and the alternatives (respectively, true and false nulls). For example, in the analysis of microarray data, it might be hard to estimate in advance the chances for each gene to be differentially expressed but typically it is believed that the number of such genes is small. A configuration of true and false null hypotheses is uniquely determined by an $n$-dimensional indicator vector $\boldsymbol{x}$, where $x_i = I\{\boldsymbol{\theta}_i \in \bar{\Theta}_i\},\ i = 1, ..., n$. Let $k = x_1 + ... + x_n$ be the number of hypotheses coming from the alternatives and assume some prior distribution $k \sim \pi(k) > 0,\ k = 0, ..., n$. Several possible choices for $\pi(k)$ are discussed later in the paper. For a given $k$, assume all $\binom{n}{k}$ various configurations of true and false null hypotheses to be equally likely *a priori*, i.e., conditionally on $k$,

$$P\left(\boldsymbol{x} \,\middle|\, \sum_{i=1}^{n} x_i = k\right) = \binom{n}{k}^{-1}. \tag{2.1}$$

To complete the prior, assume

$$(\boldsymbol{\theta}_i|x_i = 0) \sim p_{0i}(\boldsymbol{\theta}_i) \quad \text{and} \quad (\boldsymbol{\theta}_i|x_i = 1) \sim p_{1i}(\boldsymbol{\theta}_i) \tag{2.2}$$

for some specified densities $p_{0i}(\boldsymbol{\theta}_i)$ and $p_{1i}(\boldsymbol{\theta}_i)$ on $\Theta_i$ and $\bar{\Theta}_i$ respectively.

*2.2. MAP multiple testing procedure.* The Bayesian inference in multiple hypothesis testing is based on the posterior joint distribution of null hypotheses that is uniquely defined by the posterior distribution $\pi(\boldsymbol{x}, k|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$.

For the independent data $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ and the proposed hierarchical prior model, we have

$$\begin{aligned}
&\pi(\boldsymbol{x}, k|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \\
&\propto \binom{n}{k}^{-1} \pi(k) I\left\{\sum_{i=1}^{n} x_i = k\right\} \\
&\times \prod_{i=1}^{n} \left(\int_{\bar{\Theta}_i} f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i) p_{1i}(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i\right)^{x_i} \left(\int_{\Theta_i} f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i) p_{0i}(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i\right)^{1-x_i} \\
&\propto \binom{n}{k}^{-1} \pi(k) I\left\{\sum_{i=1}^{n} x_i = k\right\} \prod_{i=1}^{n} \left(\frac{\int_{\bar{\Theta}_i} f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i) p_{1i}(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}{\int_{\Theta_i} f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i) p_{0i}(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}\right)^{x_i} \\
&= \binom{n}{k}^{-1} \pi(k) I\left\{\sum_{i=1}^{n} x_i = k\right\} \prod_{i=1}^{n} (B_i^{-1})^{x_i}, \qquad (2.3)
\end{aligned}$$

where $B_i$ is the Bayes factor of $H_{0i}$ (e.g., Berger, 1985, Section 4.3.3):

$$B_i = \frac{\int_{\Theta_i} f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i) p_{0i}(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}{\int_{\bar{\Theta}_i} f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i) p_{1i}(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}. \qquad (2.4)$$

Given the posterior distribution $\pi(\boldsymbol{x}, k|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, a common approach is to select the most likely configuration of true and false null hypotheses, that is the posterior mode of (2.3). Such a Bayesian rule is also known as a maximum *a posteriori* (MAP) rule and corresponds to the (non-additive) "nothing or everything" multiple "0-1" loss discussed in the Introduction. Generally, to find the posterior mode of $\pi(\boldsymbol{x}, k|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, one should look through all $2^n$ various configurations of true and false null hypotheses. However, for the model at hand, the number of possible candidates is essentially reduced to $n+1$ only. To see this, let $\hat{\boldsymbol{x}}(k)$ be a maximizer of (2.3) for a *fixed* $k$ that indicates the most likely configuration with $k$ false null hypotheses. From (2.3), the obvious solution for $\hat{\boldsymbol{x}}(k)$ is $\hat{x}_i(k) = 1$ for the $k$ tests with the smallest Bayes factors $B_i$ and zeroes for others. Thus, one needs to choose only from $n+1$ configurations $(\hat{\boldsymbol{x}}(k), k) = H_{1(1)} \bigcap \ldots \bigcap H_{1(k)} \bigcap H_{0(k+1)} \ldots \bigcap H_{0(n)}$, $k = 0, \ldots, n$, where $H_{0(i)}$ and $H_{1(i)}$ are respectively the null and alternative hypotheses corresponding to $B_{(i)}$, $B_{(1)} \leq \ldots \leq B_{(n)}$. This leads to the following Bayesian MAP multiple testing procedure :

1. Calculate Bayes factors $B_i$, $i = 1, \ldots, n$ for all individual tests and order them in increasing sequence $B_{(1)} \leq \cdots \leq B_{(n)}$.

2. Find $\hat{k}$ that maximizes

$$\hat{\pi}_k = \pi(\hat{\boldsymbol{x}}(k), k|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \propto \binom{n}{k}^{-1} \pi(k) \prod_{i=0}^{k} B_{(i)}^{-1}, \qquad (2.5)$$

where we set $B_{(0)} = 1$ to include $k = 0$.

3. Accept all null hypotheses if $\hat{k} = 0$; otherwise, reject $\hat{k}$ null hypotheses corresponding to $B_{(1)}, \ldots, B_{(\hat{k})}$ and accept others.

One can also consider stepwise versions of the above MAP procedure. In addition to the global maximum $\hat{k}$ in (2.5), let $\hat{k}_l$ and $\hat{k}_r$ be respectively the leftmost and the rightmost local maxima of $\hat{\pi}_k$. Obviously, $\hat{k}_l \leq \hat{k} \leq \hat{k}_r$. Consider the successive ratios

$$\frac{\hat{\pi}_k}{\hat{\pi}_{k-1}} = \frac{k}{n-k+1} \frac{\pi(k)}{\pi(k-1)} B_{(k)}^{-1}.$$

Evidently, $\hat{k}_r$ is the maximal $k$ for which $\hat{\pi}_k/\hat{\pi}_{k-1} > 1$. Hence, choosing $k = \hat{k}_r$ corresponds to the Bayesian step-up (MAP-up) multiple testing procedure starting with the least significant null hypothesis with the largest Bayes factor $B_{(n)}$ and accepting null hypotheses as long as

$$B_{(k)} > \frac{k}{n-k+1} \frac{\pi(k)}{\pi(k-1)}. \qquad (2.6)$$

Analogously, $k = \hat{k}_l$ yields the Bayesian step-down (MAP-down) procedure that starts with $B_{(1)}$ and rejects null hypotheses until for the first time (2.6) holds.

## 3 Some Parallels with Existing Multiple Testing Procedures

In this Section, we discuss the relations between the presented MAP multiple testing procedures with several other existing Bayesian and frequentist procedures.

*3.1. Bayesian procedures.* Bayesians usually treat simultaneous testing of $n$ hypotheses within model selection framework, where configuration of true and false null hypotheses is considered as a disjoint partition of the parameter space $\Omega_1 \oplus \ldots \oplus \Omega_n$ (e.g., Kass and Raftery, 1995; Berger and

Pericchi, 1996). The main obstacle of this approach is that one then generally needs to compare $2^n$ possible configurations. The proposed MAP multiple testing procedure can also be viewed within such a framework but, as we have already mentioned, assuming independence of $f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i)$, there are only $n+1$ candidates to look through in our case. Similar to frequentist analysis, stepwise multiple testing procedures allow further reduction of the computational cost of the problem for large $n$. To the best of our knowledge, the only attempt to develop a Bayesian multiple stepwise testing procedure was in a recent paper of Sarkar and Chen (2004) although their hierarchical prior model is different. Using our notations, the motivation behind their procedure can be described as follows. Let $\pi_k$ and $\hat{\pi}_k$ be a prior and the corresponding posterior probability respectively of a configuration $(\hat{\boldsymbol{x}}(k), k)$. In particular, for our model $\pi_k = \binom{n}{k}^{-1}\pi(k)$ and $\hat{\pi}_k$ is given in (2.5). The *stepwise* Bayes factor $B^{(k)}$ of Sarkar and Chen (2004) is the Bayes factor of $(\hat{\boldsymbol{x}}(k), k)$ to *any* of $(\hat{\boldsymbol{x}}(i), i)$, $i = k+1, \ldots, n$ and is defined as

$$\begin{aligned} B^{(k)} &= \frac{P(\hat{\boldsymbol{x}}(k), k|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)}{P(\bigcup_{i=k+1}^n \{(\hat{\boldsymbol{x}}(i), i)|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\})} \cdot \frac{P(\bigcup_{i=k+1}^n \{(\hat{\boldsymbol{x}}(i), i)\})}{P(\hat{\boldsymbol{x}}(k), k)} \\ &= \frac{\hat{\pi}_k}{\sum_{i=k+1}^n \hat{\pi}_i} \cdot \frac{\sum_{i=k+1}^n \pi_i}{\pi_k} \ . \end{aligned}$$

One then starts from $k = 0$ and stops the first time $B^{(k)} > 1$. The main difference between such a step-down procedure and the one proposed in Section 2.2. is that the latter in effect compares $(\hat{\boldsymbol{x}}(k), k)$ against $(\hat{\boldsymbol{x}}(k+1), k+1)$ rather than against $\bigcup_{i=k+1}^n (\hat{\boldsymbol{x}}(i), i)$. The critical values for the corresponding Bayes factors also differ.

*3.2. Frequentist procedures.* As we have mentioned in the Introduction, Bayesian multiple testing procedures and their frequentist counterparts are conceptually different. The latter work with a series of $p$-values $p_i$ while the former are typically based on a sequence of Bayes factors $B_i$. O'Hagan (1995) and Berger and Pericchi (1996) proposed modifications of a standard Bayes factor for the case of improper noninformative priors but these issues are beyond the scope of the paper. The dispute about whether a Bayes factor or a $p$-value represents the better evidence for statistical inference and attempts to "reconcile" them have a long history (see e.g., Casella and Berger, 1987; Good, 1992; Kass and Raftery, 1995; Berger *et al.* 1997; Berger, 2003; Bayarri and Berger, 2004). The differences are generally due to different philosophies behind frequentist and Bayesian approaches. Nevertheless, we believe that they are not complete strangers and may benefit from each

other. Rubin (1984) argued for the importance of frequentist analysis of Bayesian procedures for better understanding and validating their results. Similarly, Bayesian interpretation of a frequentist procedure is often helpful in providing intuition behind it.

Generally, there are no immediate connections between $B_i$ and $p_i$, and direct analogies between the two approaches to hypothesis testing can be verified in some particular cases only. Consider the following, which is probably the most known case. Let $Y_i$ have independent symmetric location distributions $f_i(|y_i - \theta_i|)$, $i = 1, \ldots, n$ that have monotone likelihood ratios. Consider a set of $n$ one-sided simultaneous tests $H_{0i} : \theta_i \leq \theta_{0i}$ against $H_{1i} : \theta_i > \theta_{0i}$. Assume some prior $\pi(k)$ on the number of false $H_{0i}$, (2.1) and noninformative priors $p_{0i}(\theta_i) = 1_{(-\infty,\theta_{0i})}(\theta_i)$ and $p_{1i}(\theta_i) = 1_{(\theta_{0i},\infty)}(\theta_i)$ in (2.2). Simple straightforward calculus shows that in this case, due to the symmetry of $f_i$, $P(\theta_i \leq \theta_{0i}|Y_i) = p_i$, where $p_i$ is the $p$-value for the likelihood ratio test (e.g., Berger, 1985, Section 4.3.3). The resulting ordered Bayes factors (2.4) are evidently $B_{(i)} = p_{(i)}/(1 - p_{(i)})$, and the stepwise MAP multiple testing procedures introduced in Section 2.2. then compare $p_{(i)}$ with a sequence of critical values $p_i^* = c_i/(1+c_i)$, where $c_i$ is given in the right-hand side of (2.6). Using the relation $B_i = p_i/(1 - p_i)$ and re-writing (2.6) in terms of $p_i$, it is possible to find corresponding priors $\pi(k)$ to mimic various stepwise frequentist one-sided testing procedures. One is, however, naturally looking for some "meaningful" priors. As an illustration, consider two particular choices for $\pi(k)$.

Let $k \sim B(n, \alpha_n)$. This binomial prior suggests that each null hypothesis independently has the same prior probability $\alpha_n$ (depending possibly on $n$) of being false. Small $\alpha_n$ reflects a *sparsity* assumption that there is only a small fraction of hypotheses coming from alternatives (true alternatives or, equivalently, false nulls) whereas the majority of true hypotheses arise from nulls (true nulls). The binomial prior yields independent priors on $\boldsymbol{\theta}_i$ and the corresponding $p_i^* = \alpha_n$ are the same for all $p_{(i)}$ (see (2.6)). Obviously, $\hat{k}_l = \hat{k} = \hat{k}_r$ in this case, both stepwise versions coincide with the original MAP procedure and reject all nulls with $p$-values less than $\alpha_n$. The binomial prior, therefore, corresponds to independent testing of each individual null hypothesis at the same significance level $\alpha_n$ that implies the expected number of erroneously rejected null hypotheses $n_0\alpha_n$, where $n_0$ is the (unknown) number of true nulls. For a *super*-sparse case $E(k) = n\alpha_n < 1$, and it is equivalent to the Bonferroni procedure with the significance level $\alpha = n\alpha_n$.

For the truncated geometric prior $G_n^*(1-q)$, where $\pi(k) = (1-q)q^k/(1-q^{n+1}),\ k = 0, \ldots, n;\ 0 < q < 1$, the ordered $p_{(i)}$'s are successively compared with

$$p_i^* = \frac{i}{n - i(1-q) + 1}\, q, \qquad 1 \le i \le h,$$

which coincide with the critical values from the adaptive step-down procedure of Benjamini *et al.* (2006). The authors gave the motivation for such a procedure and reported on simulation results showing that its FDR level does not exceed $q$ although no rigourous proof was given.

The priors $\pi(k)$ corresponding to both Bonferroni and adaptive FDR procedures are sparse. In fact, this is not so surprising. The "nothing or everything" multiple "0-1" loss behind the MAP testing procedure places equal penalties for a wrong decision in both directions. On the other hand, traditional frequentist criteria in multiple testing, like FWE and FDR, mainly concern erroneous rejections of null hypotheses. Thus, both approaches are likely to yield similar results when a Bayesian believes *a priori* that the number of hypotheses arising from alternatives is small, and the total loss for their wrong rejections is, therefore, relatively low. Finally, note that for both cases, the resulting FWE and FDR levels respectively for the MAP procedure are defined by the parameters of the prior $\pi(k)$ instead of being fixed in advance at some standard level (e.g., .01 or .05). In practice, these parameters are rarely known *a priori* and should be estimated from the data (see Section 5.1.), which makes the corresponding error levels *data-adaptive*.

## 4 Normal Data

In this Section, we discuss the proposed MAP multiple testing procedure and its stepwise versions for the normal data. Let $Y_{i1}, \ldots, Y_{im_i},\ i = 1, \ldots, n$ be independent random samples of sizes $m_i$ from $N(\mu_i, \sigma_i^2)$, where the variance $\sigma_i^2$ is assumed to be known. Testing one-sided hypotheses $H_{0i} : \mu_i \le \mu_{0i}$ against $H_{1i} : \mu_i > \mu_{0i}$ is essentially a particular case of a general one-sided testing problem considered in Section 3.2, and all the results developed there can be applied directly to normal data. In this section, we focus on multiple testing of *point* null hypotheses $H_{0i} : \mu_i = \mu_{0i}$, which is probably more relevant in practice, against the two-sided alternative $H_{1i} : \mu_i \ne \mu_{0i}$. Unlike one-sided testing, frequentist and Bayesian inferences are different in this case. There are several important issues in imposing priors for point null hypotheses. To test a point null hypothesis one cannot use a continuous prior density $\pi_i(\mu_i)$ on $\mu_i$ since it yields a zero prior (and, hence, posterior)

probability for $\mu_i = \mu_{0i}$. A usual approach in this case is to set $p_{0i}$ to be a probability atom $\delta(\mu_{0i})$ at $\mu_{0i}$ and a density $p_{1i}(\mu_i)$ for $\mu_i \neq \mu_{0i}$ in (2.2), where, unlike the case of one-sided hypotheses, $p_{1i}(\mu_i)$ should be a *proper* density to avoid Lindley's paradox (e.g., Berger, 1985, Section 4.3.3). A simple and common choice is a conjugate prior $N(\mu_{0i}, \tau^2)$ (see e.g., Berger *et al.* 1997). Then straightforward calculus implies that

$$B_i = \sqrt{1+\gamma_i}\ \exp\left\{-\frac{Z_i^2}{2(1+1/\gamma_i)}\right\}, \tag{4.1}$$

where $Z_i = \sqrt{m_i}(\bar{Y}_i - \mu_{0i})/\sigma_i$ is the standardized sample average and $\gamma_i = m_i\tau^2/\sigma_i^2$ is the variance ratio (see also (4.15)–(4.17) of Berger, 1985). On the other hand, the corresponding $p$-value for the $i$-th test is $p_i = 2\tilde{\Phi}(|Z_i|)$, where $\tilde{\Phi}(\cdot) = 1-\Phi(\cdot)$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. From (4.1), one has $B_i \leq \sqrt{1+\gamma_i}$ and

$$p_i = 2\tilde{\Phi}\left(\left(2(1+1/\gamma_i)\ln\left\{\sqrt{1+\gamma_i}/B_i\right\}\right)^{1/2}\right), \tag{4.2}$$

which is an increasing function of $B_i$. Hence, imposing a prior $\pi(k)$ on the number of false null hypotheses, (2.6) in the stepwise MAP multiple testing procedures becomes equivalent to comparing $p_{(i)}$ with the critical value $p_i^*$, where

$$p_i^* = 2\tilde{\Phi}\left(\left(2(1+1/\gamma_i)\ln_+\left\{\sqrt{1+\gamma_i}\cdot\frac{n-i+1}{i}\cdot\frac{\pi(i-1)}{\pi(i)}\right\}\right)^{1/2}\right), \tag{4.3}$$

and $\ln_+(x) = \max(0, \ln(x))$.

Consider several choices for $\pi(k)$ with the same mean $\lambda_n$ depending possibly on $n$. For simplicity of exposition, assume that all samples have the same variance $\sigma^2$ and are of an equal size $m$ so that $\gamma_i = \gamma = m\tau^2/\sigma^2$. For the binomial prior $B(n, \xi_n)$, where $\xi_n = \lambda_n/n$, (4.3) yields the same critical value for all tests

$$p^* = 2\tilde{\Phi}\left(\left(2(1+1/\gamma)\ln_+\left\{\sqrt{1+\gamma}\,\frac{1-\xi_n}{\xi_n}\right\}\right)^{1/2}\right). \tag{4.4}$$

From a frequentist viewpoint, it corresponds to testing each individual null hypothesis at the significance level $p^*$ and, therefore, controlling the expected number of erroneously rejected null hypotheses at the level $n_0 p^*$. As in Section 3.2., $p^*$ is defined by the parameters of the prior and the noise level

instead of being fixed in advance at .01 or .05, and, in practice, it is typically estimated from the data (see Section 5.1.). Assuming $\lambda_n = o(n)$ (sparsity) for sufficiently large $m$, one gets a simpler approximation $p_a^*$ for $p^*$ :

$$\begin{aligned} p_a^* &\sim 2\tilde{\Phi}\left(\left(2\ln\left\{\sqrt{\gamma}\,\frac{1-\xi_n}{\xi_n}\right\}\right)^{1/2}\right) \\ &\sim \frac{\xi_n}{1-\xi_n}\left(\pi\gamma\ln\left\{\sqrt{\gamma}\frac{1-\xi_n}{\xi_n}\right\}\right)^{-1/2} \\ &\sim \frac{\xi_n}{\sqrt{\pi\gamma\ln(\sqrt{\gamma}/\xi_n)}}, \end{aligned} \tag{4.5}$$

where we exploited the well known asymptotic relation $\tilde{\Phi}(x) \sim \phi(x)/x$ for large $x$, where $\phi(\cdot)$ is the standard normal density (e.g., Barndorff-Nielsen and Cox, 1989, p.56). If, in addition, $\lambda_n < \sqrt{\pi\gamma\ln n}$ (super-sparse case), it follows immediately that the MAP multiple testing procedure with the binomial prior is a Bonferroni procedure with the FWE controlling level

$$\alpha_n = np^* \sim \frac{\lambda_n}{\sqrt{\pi\gamma\ln(\sqrt{\gamma}/\xi_n)}} \quad (<1) \tag{4.6}$$

(see (4.5)).

Consider now the truncated Poisson distribution $Pois^*(\lambda_n)$, where

$$\pi(k) = \frac{\lambda_n^k/k!}{\sum_{j=0}^n \lambda_n^j/j!},\ k = 0,\dots,n. \tag{4.7}$$

Lemma 4.1. *Suppose $k$ has the truncated Poisson distribution $Pois^*(\lambda_n)$ (4.7), where $\lambda_n = o(n)$. Then,*

*(i) $E(k) = \lambda_n(1-\delta_n)$, where a positive sequence $\delta_n = o(1)$, and*

*(ii) $k = o(n)$ almost surely.*

The proof of Lemma 4.1 is given in the Appendix. From (4.3) and Lemma 4.1, one has

$$\begin{aligned} p_i^* &= 2\tilde{\Phi}\left(\left(2(1+1/\gamma)\ln_+\left\{\sqrt{1+\gamma}\cdot\frac{n-i+1}{\lambda_n}\right\}\right)^{1/2}\right) \\ &\sim \frac{\lambda_n}{n-i+1}\cdot\frac{1}{\sqrt{\pi\gamma\ln(\sqrt{\gamma}/\xi_n)}}. \end{aligned}$$

Let again $\lambda_n < \sqrt{\pi\gamma \ln n}$. The corresponding step-up and step-down procedures in this case are, in fact, the FWE controlling procedures of Hochberg (1988) and Holm (1979) respectively with the same FWE level $\alpha_n$ in (4.6). Therefore, the original MAP procedure sandwiched between its two stepwise versions also asymptotically controls the FWE at the level $\alpha_n$.

Consider also a "reflected" truncated Poisson distribution, where

$$\pi(k) = \frac{(n-\lambda_n)^{n-k}/(n-k)!}{\sum_{j=0}^{n}(n-\lambda_n)^{n-j}/(n-j)!},\ k = 0, \ldots, n. \tag{4.8}$$

LEMMA 4.2. *Let $\pi(k)$ be given by (4.8), where $\lambda_n = o(n)$ but $\lambda_n/\sqrt{n \ln n} \to \infty$. Then,*

*(i) $E(k) = \lambda_n(1+\delta_n)$, where a positive sequence $\delta_n = o(1)$, and*

*(ii) $k = \lambda_n(1+o(1))$ almost surely.*

The proof of Lemma 4.2 is given in the Appendix. Lemma 4.2 and (4.3) imply

$$\begin{aligned} p_i^* &= 2\tilde{\Phi}\left(\left(2(1+1/\gamma)\ln_+\left\{\sqrt{1+\gamma}\cdot\frac{n-\lambda_n}{i}\right\}\right)^{1/2}\right) \\ &\sim \frac{i}{n-\lambda_n}\cdot\frac{1}{\sqrt{\pi\gamma\ln(\sqrt{\gamma}/\xi_n)}} = \frac{i}{n}\cdot q_n, \end{aligned} \tag{4.9}$$

where $q_n = (1-\xi_n)^{-1}(\pi\gamma\ln(\sqrt{\gamma}/\xi_n))^{-1/2}$. Within a frequentist framework, the corresponding step-up procedure mimics the well-known step-up FDR controlling procedure of Benjamini and Hochberg (1995) with the FDR parameter $q_n$. They proved that its FDR level does not exceed $(n_0/n)q_n \le q_n$. Later, Benjamini and Yekutieli (2001) established that the FDR level is, in fact, exactly $(n_0/n)q_n$. Sarkar (2002) showed that it is also true for the corresponding step-down procedure. In addition, Lemma 4.2 indicates that for sparse cases typically $\hat{k}_l \sim \hat{k} \sim \hat{k}_r \sim \lambda\ (\le E(k))$ and all three versions of the MAP testing procedure yield similar results (see also Section 5). The above examples indicate again that the traditional frequentist procedures focusing mainly on the control of erroneous rejections of null hypotheses, correspond to *sparse* priors, where the expected proportion of false null hypotheses tends to zero. For extremely conservative FWE controlling procedures, the expected number of false null hypotheses grows at most at logarithmic rate (*super-sparse* priors), while for less stringent FDR controlling procedures it grows as $n^\beta$, $0 < \beta < 1$. We have discussed the reasons for such sparsity of priors in Section 3.2.

## 5 Simulated Example

To test the performance of the proposed procedure, we considered a simulated example, where the data was generated according to the normal model considered in Section 4 :

$$Y_{i,j} = \mu_i + \epsilon_{i,j} \qquad i = 1, \ldots, n;\ j = 1, \ldots, m, \tag{5.1}$$

where the $\epsilon_{i,j}$'s are i.i.d. $N(0, \sigma^2)$ variables. Such an example is similar to that of Ishwaran and Rao (2003) motivated by the analysis of microarray data. In microarray experiments, expression levels of thousands of genes present in a biological sample are simultaneously measured to identify a small proportion of differentially expressed genes. The model (5.1) is supposed to mimic an experiment where two groups of biological samples (control and treatment) are marked using two dyes (usually Cy5 and Cy3 for red and green respectively) and then hybridized on several cDNA microarrays. In such experimental context, the value $Y_{i,j}$ represents the log intensity ratio between the control and the treatment groups for the $i$-th gene on the $j$-th array after a suitable normalization procedure. To detect differentially expressed genes, one simultaneously tests $H_{0i} : \mu_i = 0$ against $H_{1i} : \mu_i \neq 0$ for $i = 1, \ldots, n$.

Let $\pi(k)$ be a specified prior distribution on the number of differentially expressed genes with $E(k) = \lambda$. Given $k$, all $\binom{n}{k}$ possible configurations were assumed to be equally likely. Unexpressed genes naturally had a zero expression level while the expression levels $\mu_i$ of differentially expressed genes were simulated as independent $N(0, \tau^2)$ variables. A similar simulation model was considered in Ishwaran and Rao (2003) but with the same *fixed* $\mu$ for all false nulls. We believe that our model is more realistic for the microarray data where gene expression levels may be different.

It should be noted that model (5.1) might be, in fact, somewhat idealized for real microarray data. In particular, it ignores co-regulations between gene expressions. Storey and Tibshirani (2001) argue that these co-regulations are commonly of "clumpy" type, where different genes can be clustered in small independent groups having within group correlations. Several simulation studies (see e.g. Storey and Tibshirani, 2001; Ishwaran and Rao, 2003; Reiner, Yekutieli and Benjamini, 2003) indicate though that clumpy dependencies have a relatively minor effect when group sizes are small compared to the total number of tested hypotheses — a very reasonable scenario with microarray data. In addition, the assumption of equal variances is not always valid. The proposed MAP testing procedure can still be adapted for

different $\sigma_i^2$ (see Section 4) but estimation of the parameters of the model described in Section 5.1 below should be modified in this case. Alternatively, sometimes the data can be transformed first through a variance-stabilization transformation (e.g. logarithmic).

*5.1. Estimation of parameters.* To apply the developed MAP procedures for a chosen $\pi(k)$, one still needs to specify the noise variance $\sigma^2$ and the prior variance $\tau^2$ or, equivalently, the variance ratio $\gamma = m\tau^2/\sigma^2$. These parameters are rarely known *a priori* in practice and should be estimated from the data in the spirit of empirical Bayes. Moreover, we also assume that $\pi(k)$ has a prescribed parametric form with an unknown expectation $\lambda$ that should be estimated as well.

Straightforward calculus yields the following marginal likelihood of the observed data $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y_n})'$, $i = 1, \ldots, n$, where $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,m})'$:

$$L(\boldsymbol{Y}; \lambda, \sigma^2, \gamma) = \sum_{k=0}^{n} \pi(k) \binom{n}{k}^{-1} \sum_{\boldsymbol{x}_j : \sum_i x_{ji} = k} \prod_{i=1}^{n} f_1(\boldsymbol{Y}_i)^{x_{ji}} f_2(\boldsymbol{Y}_i)^{1-x_{ji}},$$

where $\boldsymbol{x}_j$ is the indicator vector, and $f_1$ and $f_2$ are $m$-variate normal densities with zero means and variance matrices $V = \sigma^2(I_m + (\gamma/m)\mathbf{1}\mathbf{1}')$ and $\sigma^2 I_m$ respectively. Following the ideas of empirical Bayes approach, we estimate the unknown parameters $\lambda$, $\sigma^2$ and $\gamma$ by the corresponding marginal maximum likelihood estimators (MLEs). There are no closed form solutions and the EM algorithm is applied to obtain the MLEs numerically.

Regard the indicator vector $\boldsymbol{x}$ as a "missing" data. The complete log-likelihood for the "augmented" data $(\boldsymbol{Y}, \boldsymbol{x})$ is then

$$l(\boldsymbol{Y}, \boldsymbol{x}; \lambda, \sigma^2, \gamma) = \ln \pi(k) - \ln \binom{n}{k} + \sum_{i=1}^{n} x_i \ln f_1(\boldsymbol{Y}_i) + \sum_{i=1}^{n} (1 - x_i) \ln f_2(\boldsymbol{Y}_i),$$

where $k = x_1 + \ldots + x_n$.

At the $h$-th iteration, the E-step consists of computing the conditional expectation

$$\begin{aligned} \hat{l}_{[h]} &= E\left( l(\boldsymbol{Y}, x) | \boldsymbol{Y}, \lambda_{[h]}, \sigma^2_{[h]}, \gamma_{[h]} \right) \\ &= E\left( \ln \left( \pi(k) \binom{n}{k}^{-1} \right) \,\middle|\, \boldsymbol{Y}, \lambda_{[h]} \right) + \sum \eta_{i[h]} \ln f_1(\boldsymbol{Y}_i) \\ &\quad + \sum_{i=1}^{n} \left(1 - \eta_{i[h]}\right) \ln f_2\left(\boldsymbol{Y}_i\right), \end{aligned}$$

where $\eta_{i[h]} = E\left(x_i|\boldsymbol{Y}_i, \lambda_{[h]}, \sigma^2_{[h]}, \gamma_{[h]}\right) = \left(1 + B_{i[h]}\frac{n-\lambda_{[h]}}{\lambda_{[h]}}\right)^{-1}$ and the Bayes factor $B_{i[h]} = \left(1+\gamma_{[h]}\right)^{1/2} \exp\left\{-m\bar{Y}_i^2/\left(2\sigma^2_{[h]}\left(1+1/\gamma_{[h]}\right)\right)\right\}$ (cf. (4.1)).

Straightforward calculus shows that, regardless of the prior $\pi(k)$, maximizing $\hat{l}_{[h]}$ with respect to $\hat{\sigma}^2_{[h+1]}$ and $\hat{\gamma}_{[h+1]}$ under the positivity constraints on the M-step results in the following solutions :

$$\begin{aligned}
\hat{\gamma}_{[h+1]} &= \max\left(0, \quad \frac{m\sum_{i=1}^n \eta_{i[h]}\bar{Y}_i^2}{\hat{\sigma}^2_{0[h+1]}\sum_{i=1}^n \eta_{i[h]}} - 1\right), \\
\hat{\sigma}^2_{[h+1]} &= \begin{cases} \hat{\sigma}^2_{0[h+1]} & \text{if } \hat{\gamma}_{[h+1]} > 0 \\ \sum_{i=1}^n\sum_{j=1}^m Y_{i,j}^2/(nm) & \text{if } \hat{\gamma}_{[h+1]} = 0, \end{cases}
\end{aligned}$$

where

$$\hat{\sigma}^2_{0[h+1]} = \frac{\sum_{i=1}^n \eta_{i[h]}\sum_{j=1}^m (Y_{i,j}-\bar{Y}_i)^2 + \sum_{i=1}^n\left(1-\eta_{i[h]}\right)\sum_{j=1}^m Y_{i,j}^2}{nm - \sum_{i=1}^n \eta_{i[h]}}.$$

The solution for $\hat{\lambda}_{[h+1]}$ is a maximizer of

$$m(\lambda) = E\left\{\ln\left(\pi(k)\binom{n}{k}^{-1}\right) \;\middle|\; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \lambda\right\}$$

and it depends on the particular type of $\pi(k)$. Consider the three priors $\pi(k)$ discussed in the Section 4. For the binomial prior $B(n, \lambda/n)$, one has

$$\begin{aligned}
m(\lambda) &= E(k|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \lambda)\ln\frac{\lambda}{n-\lambda} + n\ln\frac{n-\lambda}{n} \\
&= \sum_{i=1}^n \eta_i \ln\frac{\lambda}{n-\lambda} + n\ln\frac{n-\lambda}{n},
\end{aligned}$$

and solving the equation $m'(\lambda) = 0$ immediately yields $\hat{\lambda}_{[h+1]} = \sum_{i=1}^n \eta_{i[h]}$.

For the truncated Poisson prior $Pois^*(\lambda)$ in (4.7),

$$m'(\lambda) = \frac{\sum_{i=1}^n \eta_i - \lambda}{\lambda} + \frac{\lambda^n/n!}{\sum_{j=0}^n \lambda^j/j!}.$$

After finding the root of $m'(\lambda)$ under the conditions of Lemma 4.1, one has $\hat{\lambda}_{[h+1]} = \sum_{i=1}^n \eta_{i[h]}(1+o(1)) \sim \sum_{i=1}^n \eta_{i[h]}$.

Similarly, for the reflected truncated Poisson prior (4.8),

$$m'(\lambda) = \frac{\sum_{i=1}^{n} \eta_i - \lambda}{n - \lambda} + \frac{(n-\lambda)^n/n!}{\sum_{j=0}^{n}(n-\lambda)^j/(n-j)!}$$

and, under the conditions of Lemma 4.2, again $\hat{\lambda}_{[h+1]} = \sum_{i=1}^{n} \eta_{i[h]}(1+o(1)) \sim \sum_{i=1}^{n} \eta_{i[h]}$. Hence, for all the above three priors $\pi(k)$, the EM algorithm results in essentially the same MLEs for $\gamma$, $\sigma^2$ and $\lambda$.

*5.2. The results.* Data were generated according to the model (5.1), where the $\mu_i$'s were randomly sampled from $N(0, \tau^2)$ in $\xi$ percent of the cases, and set to zero in the remaining cases. This corresponds to a scenario where only $\xi\%$ of genes are differentially expressed. We set $\sigma = 1$ and $\tau = 2$. To mimic a typical microarray experiment, we choose a large number of tested hypotheses (genes) $n = 10000$ with a small proportion $\xi = .05$ of false nulls (differentially expressed genes) corresponding to a sparse (though not super-sparse) case. Several values of $m$ were also tried. The number of replications was 1000.

The true values of $\sigma^2$, $\tau^2$ and $\xi$ were assumed to be unknown in simulations and were estimated by the EM-algorithm described in the previous Section 5.1. We investigated two MAP multiple testing procedures corresponding to the binomial prior $B(n, \xi)$ (Binomial) and the reflected truncated Poisson prior (4.8), where $\pi(k) \propto (n - n\xi)^{n-k}/(n-k)!$ (Pois2). For both priors $\lambda = E(k) \sim n\xi$ and the EM-estimates for the parameters are the same (see Section 5.1.). The entire study was carried out using the MATLAB programming environment.

Table 1 summarizes the EM estimates of $\xi$, $\sigma^2$ and $\gamma = m\tau^2/\sigma^2$ averaged over 1000 replications and their standard deviations. As expected, for the given variance ratio $\tau^2/\sigma^2$, the accuracy of estimation improves as $\gamma$ increases but even for moderate $\gamma$ it is quite satisfactory. Similar results were obtained for $\xi = .01, .1, .2, .3$.

TABLE 1. MEANS AND STANDARD DEVIATIONS FOR THE EM-ESTIMATES OF THE UNKNOWN PARAMETERS $\xi$, $\sigma^2$ AND $\gamma$ AVERAGED OVER 1000 REPLICATIONS. THE TRUE VALUES ARE $\xi = .05$, $\sigma = 1$ AND $\tau^2 = 4$.

| $m$ | $\gamma$ | $\hat{\xi}$ | $\hat{\sigma}^2$ | $\hat{\gamma}$ |
|---|---|---|---|---|
| 5 | 20 | .04859 (.00261) | 1.00060 (.00657) | 21.60540 (1.66995) |
| 10 | 40 | .04960 (.00194) | 1.00030 (.00457) | 41,37186 (2.91361) |
| 20 | 80 | .04988 (.00152) | 1.00013 (.00312) | 81.54536 (5.52843) |
| 30 | 120 | .04988 (.00138) | 1.00011 (.00257) | 121.49824 (8.52774) |

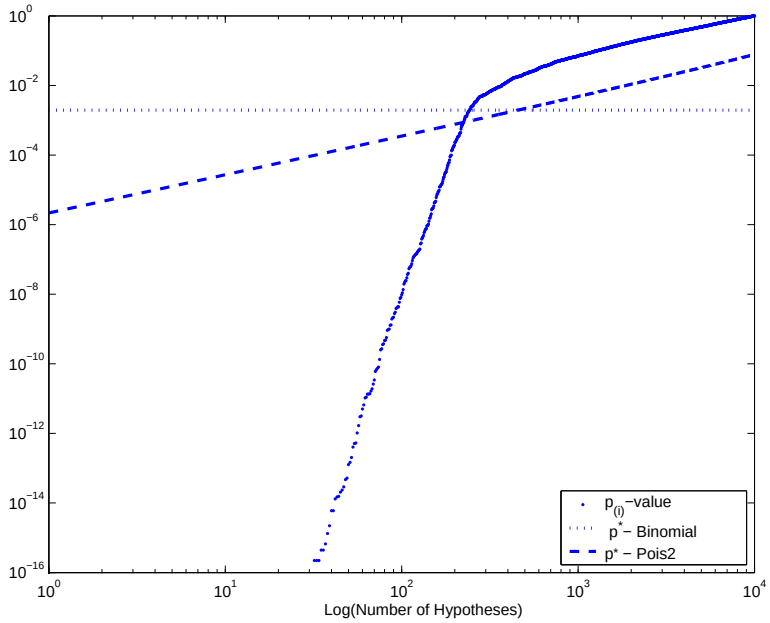


Figure 1. Ordered $p$-values $p_{(i)}$ and the corresponding critical values $p_i^*$ for Binomial and Pois2 on the log-log scale for a single realization from (5.1) with $n = 10000$, $m = 5$, $\sigma = 1$ and $\xi_n = .05$.

Figure 1 shows the ordered $p$-values as in (4.2) and the corresponding critical values $p_i^*$ as in (4.4) and (4.9) for Binomial and Pois2 priors respectively on the log-log scale evaluated from a single realization from (5.1). In most realizations, there was a single intersection point for Pois2 prior, and therefore all the three versions of the MAP procedure were identical.

The main simulation results on the performance of MAP multiple testing procedures are presented in Table 2 and Table 3 for $m = 5$ (small samples) and $m = 20$ (fairly large samples) respectively. The tables give the total number of detections (rejected nulls), the total number of erroneous decisions (misclassifications), the proportions of erroneous rejections of nulls and alternatives (Type I and Type II errors), and the proportions of erroneously rejected null hypotheses among those being rejected (false discovery rates, FDR) and erroneously nonrejected null hypotheses among those being accepted (false negative rates, FNR) for the original MAP procedures and for their stepwise versions. They are also compared with the two well-known multiple testing procedures: the Bonferroni procedure with familywise error levels $\alpha = .01$ and .05 and the step-up FDR controlling procedure of Benjamini and Hochberg (1995) with the FDR parameters $q = .01$ (BH.01) and

.05 (BH.05). The results in Table 2 and Table 3 are averaged over 1000 replications. Standard errors are given in the brackets. The unknown parameters were estimated by the EM-algorithm.

TABLE 2. PERFORMANCES OF MAP, BONFERRONI AND BENJAMINI-HOCHBERG MULTIPLE TESTING PROCEDURES ($m = 5$, $\xi = .05$).

| Method | Detect | Miss-classified | Type I Error | Type II Error | FDR | FNR |
|---|---|---|---|---|---|---|
| Binomial | 270 | 269 | .00207 | .49874 | .07249 | .02563 |
| | (.45124) | (.38654) | (.00002) | (.00078) | (.00052) | (.00004) |
| Pois2 (step-down) | 244 | 274 | .00095 | .52938 | .03657 | .02713 |
| | (.43028) | (.39577) | (.00001) | (.00080) | (.00041) | (.00004) |
| Pois2 (step-up) | 244 | 274 | .00095 | .52935 | .03658 | .02713 |
| | (.43015) | (.39571) | (.00001) | (.00080) | (.00041) | (.00004) |
| Pois2 (global) | 244 | 274 | .00095 | .52937 | .03657 | .02713 |
| | (.43022) | (.39573) | (.00001) | (.00080) | (.00041) | (.00004) |
| Bonferroni (.05) | 160 | 341 | .00000 | .68096 | .00024 | .03460 |
| | (.33067) | (.33058) | (.00000) | (.00066) | (.00004) | (.00003) |
| Bonferroni (.01) | 143 | 357 | .00000 | .71397 | .00006 | .03622 |
| | (.31796) | (.31800) | (.00000) | (.00064) | (.00002) | (.00003) |
| BH.05 | 253 | 271 | .00125 | .51854 | .04690 | .02660 |
| | (.42965) | (.38663) | (.00001) | (.00078) | (.00044) | (.00004) |
| BH.01 | 212 | 292 | .00021 | .58081 | .00941 | .02967 |
| | (.38999) | (.38262) | (.00000) | (.00077) | (.00022) | (.00004) |

TABLE 3. PERFORMANCES OF MAP, BONFERRONI AND BENJAMINI-HOCHBERG MULTIPLE TESTING PROCEDURES ($m = 20$, $\xi = .05$).

| Method | Detect | Miss-classified | Type I Error | Type II Error | FDR | FNR |
|---|---|---|---|---|---|---|
| Binomial | 372 | 152 | .00123 | .28011 | .03131 | .01454 |
| | (.35794) | (.33236) | (.00001) | (.00065) | (.00030) | (.00003) |
| Pois2 (step-down) | 364 | 152 | .00087 | .28806 | .02270 | .01495 |
| | (.35089) | (.33121) | (.00001) | (.00065) | (.00026) | (.00003) |
| Pois2 (step-up) | 364 | 152 | .00087 | .28805 | .02271 | .01495 |
| | (.35077) | (.33098) | (.00001) | (.00065) | (.00026) | (.00003) |
| Pois2 (global) | 364 | 152 | .00087 | .28806 | .02270 | .01495 |
| | (.3584) | (.33111) | (.00001) | (.00065) | (.00026) | (.00003) |
| Bonferroni (.05) | 306 | 194 | .00001 | .38774 | .00017 | .02000 |
| | (.34086) | (.34124) | (.00000) | (.00068) | (.00002) | (.00003) |
| Bonferroni (.01) | 294 | 206 | .00000 | .41204 | .00003 | .02127 |
| | (.34688) | (.34683) | (.00000) | (.00069) | (.00001) | (.00003) |
| BH.05 | 383 | 153 | .00191 | .26946 | .04711 | .01401 |
| | (.36714) | (.33741) | (.00002) | (.00064) | (.00036) | (.00003) |
| BH.01 | 349 | 157 | .00034 | .30845 | .00928 | .01598 |
| | (.33686) | (.33830) | (.00001) | (.00066) | (.00017) | (.00003) |

The Bonferroni procedures, as expected, were extremely conservative and resulted in the lowest numbers of detections and poor power properties. To better understand the results in Table 2 and Table 3, recall that for the binomial prior, the expected Type I error of the resulting MAP procedure is $p^*$ given by (4.4). Table 4 compares Type I error from simulations with $p^*$ and its simpler approximation (for large $\gamma$ and small $\xi$) $p_a^*$ from (4.5) calculated using the true parameters. The Type I error reported in the table is the average over 1000 replications, and the unknown parameters are estimates with the procedure described in Section 5.1.. One sees that simulated Type I errors and $p^*$ match nicely. BH.01 and BH.05 controlled FDR similar to the expected levels $.95 \times .01 = .0095$ and $.95 \times .05 = .0475$ respectively. For all cases, the simulated FDR levels of Pois2 procedures were between those of BH.01 and BH.05, and that explains why the results of Pois2 procedures were sandwiched between more conservative BH.01 and less stringent BH.05. In fact, from the results of Section 4, for large $\gamma$ and small $\xi$, the step-up and step-down Pois2 should behave similar to the FDR controlling procedures with the (adaptive) FDR level of about $(n_0/n)q_n \sim 1/\sqrt{\pi\gamma \ln(\sqrt{\gamma}/\xi)}$ (see (4.9)).

Table 4. Simulated Type I errors for the Binomial prior and the corresponding $p^*$ and $p_a^*$.

| | $m = 5$ ($\gamma = 20$) | | | $m = 20$ ($\gamma = 80$) | | |
|---|---|---|---|---|---|---|
| $\xi$ | simul. | $p^*$ | $p_a^*$ | simul. | $p^*$ | $p_a^*$ |
| .01 | .00032 | .00038 | .00051 | .00021 | .00020 | .00024 |
| .05 | .00207 | .00219 | .00202 | .01230 | .00125 | .00138 |
| .1 | .00491 | .00519 | .00438 | .02850 | .00285 | .00297 |
| .2 | .01265 | .01345 | .01431 | .00699 | .00706 | .00647 |
| .3 | .02404 | .02570 | .02302 | .01285 | .01302 | .01027 |

Table 5 shows simulated FDR for the Pois2 prior and the corresponding approximated expected FDR from (4.9). The results show that for $m = 20$ ($\gamma = 80$), simulated FDR of Pois2 approaches the expected level. Binomial MAP procedure tends to reject more null hypotheses than Pois2. However, fairly large differences in Type I errors in favour of Pois2 and relatively negligible differences in Type II errors indicate that most of the additional detections made by Binomial were false. In fact, this is not surprising since for sparse (but not super-sparse) cases, Binomial essentially is not supposed to control any multiplicity criterion (FWE or FDR) while Pois2 is similar to the FDR controlling procedure. Finally, note that the performance of all the considered procedures improves with increasing $\gamma$.

TABLE 5. SIMULATED FDR FOR THE POIS2 PRIOR AND THE CORRESPONDING APPROXIMATED EXPECTED FDR FROM (4.9).

| | | $m = 5$ ($\gamma = 20$) | | $m = 20$ ($\gamma = 80$) | |
|---|---|---|---|---|---|
| $\xi$ | procedure | simul. | approx. | simul. | approx. |
| .01 | Step-Down | .02945 | .05106 | .01957 | .02419 |
| | Step-Up | .02957 | | .01957 | |
| | Global | .02948 | | .01957 | |
| .05 | Step-Down | .03657 | .05951 | .02270 | .02769 |
| | Step-Up | .03658 | | .02271 | |
| | Global | .03657 | | .02270 | |
| .1 | Step-Down | .04095 | .06471 | .02497 | .02975 |
| | Step-Up | .04097 | | .02497 | |
| | Global | .04096 | | .02497 | |
| .2 | Step-Down | .04702 | .07156 | .02719 | .03235 |
| | Step-Up | .04703 | | .02719 | |
| | Global | .04702 | | .02719 | |
| .3 | Step-Down | .05226 | .07675 | .02918 | .03423 |
| | Step-Up | .05227 | | .02918 | |
| | Global | .05227 | | .02918 | |

## 6 Concluding Remarks

The paper considered the multiple hypothesis testing within a Bayesian framework. We proposed a hierarchical prior model, where a prior distribution $\pi(k)$ is imposed on the number of hypotheses arising from alternatives, and then applied the maximum *a posteriori* (MAP) rule to find the most plausible configuration of true and false null hypotheses. In the case of independent likelihoods, the resulting MAP procedure and its closely related step-up and step-down versions are intuitively clear and computationally inexpensive. They compare ordered Bayes factors of individual hypotheses with a sequence of critical values depending on the prior and, in this sense, are similar in spirit to frequentist multiple testing procedures based on ordered $p$-values. By choosing different $\pi(k)$, one can mimic various existing frequentist testing procedures. In particular, for the normal data, the specific choices of $\pi(k)$ lead to several known FWE and FDR controlling procedures. We showed that the FDR controlling procedures are related to sparse $\pi(k)$, while their much more conservative FWE counterparts correspond to super-sparse priors. Furthermore, the resulting FWE and FDR levels for the Bayesian procedures are defined by the parameters of the prior rather than being fixed in advance at some traditional level (e.g., .01 or .05). In practice, the prior parameters are typically unknown and should be estimated from the data, which makes the corresponding error levels data-adaptive.

The Bayesian MAP testing procedures demonstrated good performance in a simulated example that mimics a microarray experiment.

The proposed general Bayesian approach can be used as a tool in a wide range of multiple hypothesis testing set-ups for various types of data though further analysis is needed for each specific problem at hand. Its use for model selection is another interesting topic for further research.

## Appendix: Proofs of Lemmas from Section 4

*6.1. Proof of Lemma 4.1.*

(i) Let $a_k = \lambda_n^k/k!$. We have

$$E(k) = \frac{\sum_{k=0}^{n} k a_k}{\sum_{k=0}^{n} a_k} = \lambda_n \frac{\sum_{k=0}^{n-1} a_k}{\sum_{k=0}^{n} a_k} = \lambda_n \left(1 - \frac{a_n}{\sum_{k=0}^{n} a_k}\right) < \lambda_n. \tag{6.1}$$

On the other hand, the ratio $a_n/\sum_{k=0}^{n} a_k$ in (6.1) satisfies

$$\frac{a_n}{\sum_{k=0}^{n} a_k} < \frac{a_n}{\max_k a_k} = \frac{a_n}{a_{\lambda_n}} = \frac{\lambda_n^n}{n!}\frac{\lambda_n!}{\lambda_n^{\lambda_n}}. \tag{6.2}$$

Exploiting the Stirling formula, for every $k$, one has

$$k^k e^{-k}\sqrt{2\pi k} < k! < k^k e^{-k+\frac{1}{12k}}\sqrt{2\pi k}, \tag{6.3}$$

and, hence, for $\lambda_n = o(n)$, (6.2) implies

$$\frac{a_n}{\sum_{k=0}^{n} a_k} < \left(\frac{\lambda_n}{n} e^{1-\lambda_n/n}\right)^n \sqrt{\frac{\lambda_n}{n}} e^{1/(12\lambda_n)} \to 0$$

as $n$ increases.

(ii) Consider an arbitrary sequence $\epsilon_n = o(n)$ and let $S_n = P(k \geq \lambda_n + \epsilon_n)$. In what follows, we will find a particular $\epsilon_n$ such that the series $\sum_n S_n$ will converge. The first Borel-Cantelli lemma will imply then that, for $\lambda_n = o(n)$, $k = o(n)$ almost surely. One has

$$\begin{aligned} S_n &= \frac{\sum_{k=\lambda_n+\epsilon_n}^{n} a_k}{\sum_{k=0}^{n} a_k} < \frac{a_{\lambda_n+\epsilon_n}(n-\lambda_n-\epsilon_n)}{a_{\lambda_n}} \\ &= \frac{\lambda_n^{\lambda_n+\epsilon_n}}{\lambda_n^{\lambda_n}} \cdot \frac{\lambda_n!(n-\lambda_n-\epsilon_n)}{(\lambda_n+\epsilon_n)!} < n\lambda_n^{\epsilon_n}\frac{\lambda_n!}{(\lambda_n+\epsilon_n)!}. \end{aligned} \tag{6.4}$$

Applying (6.3) to get respectively the upper and lower bounds for the factorials in the numerator and denominator of (6.4), straightforward calculus yields

$$S_n < n\kappa_n\, e^{\frac{1}{12\lambda_n}} < Cn\kappa_n$$

for some $C > 0$, where

$$\kappa_n = \frac{e^{\epsilon_n}}{(1+\epsilon_n/\lambda_n)^{\lambda_n+\epsilon_n+1/2}}.$$

Hence,

$$\ln \kappa_n < \epsilon_n \left(1 - \ln \frac{\epsilon_n}{\lambda_n}\right).$$

Consider $\epsilon_n = \max(\ln n, e^{\zeta}\lambda_n) = o(n)$, where $\zeta > 3$. Then,

$$S_n < Cne^{-\epsilon_n(\zeta-1)} < Cn^{-(\zeta-2)},$$

and, therefore, the series $\sum_n S_n$ converges. □

*6.2. Proof of Lemma 4.2.*

(i) Let $b_k = (n-\lambda_n)^k/k!$. After simple algebra one has

$$E(k) = \frac{\sum_{k=0}^{n}(n-k)b_k}{\sum_{k=0}^{n} b_k} = \lambda_n + \frac{(n-\lambda_n)b_n}{\sum_{k=0}^{n} b_k} > \lambda_n.$$

On the other hand, under the conditions of Lemma 4.2 on $\lambda_n$,

$$\lambda_n + \frac{(n-\lambda_n)b_n}{\sum_{k=0}^{n} b_k} < \lambda_n + \frac{n-\lambda_n}{\lambda_n} = \lambda_n\left(1 - 1/\lambda_n + n/\lambda_n^2\right) < \lambda_n(1+\delta_n),$$

where $\delta_n = o(1)$ and positive.

(ii) Consider the sequence $b_k$ defined above. For any positive sequence $\epsilon_n$ define $S_{n1} = P(k \geq \lambda_n + \epsilon_n)$ and $S_{n2} = P(k \leq \lambda_n - \epsilon_n)$. Similar to the ideas of the proof of Lemma 4.1, we will find a particular $\epsilon_n = o(\lambda_n)$ guaranteeing the convergence of both $\sum_n S_{n1}$ and $\sum_n S_{n2}$ under the conditions of Lemma 4.2, and, hence, by the first Borel-Cantelli lemma, we have $k = \lambda_n(1+o(1))$ with probability one. Note that $\max_k b_k = b_{n-\lambda_n}$, $b_k$ increases for $k \leq n-\lambda_n$ and decreases otherwise. One has

$$S_{n1} = \frac{\sum_{k=0}^{n-\lambda_n-\epsilon_n} b_k}{\sum_{k=0}^{n} b_k} < \frac{nb_{n-\lambda_n-\epsilon_n}}{b_{n-\lambda_n}} = \frac{n}{(n-\lambda_n)^{\epsilon_n}} \cdot \frac{(n-\lambda_n)!}{(n-\lambda_n-\epsilon_n)!}.$$

Applying (6.3) and straightforward calculus, one gets

$$S_{n1} < n\kappa_{n1}\sqrt{\frac{n-\lambda_n}{n-\lambda_n-\epsilon_n}}\, e^{1/(12(n-\lambda_n))}, \tag{6.5}$$

where

$$\kappa_{n1} = \left(\left(1+\frac{\epsilon}{n-\lambda_n-\epsilon_n}\right)^{(n-\lambda_n-\epsilon_n)/\epsilon_n} e^{-1}\right)^{\epsilon_n}.$$

The Taylor expansion of $\ln\kappa_{n1}$ implies

$$\ln\kappa_{n1} = -\frac{1}{2}\cdot\frac{\epsilon_n^2}{n-\lambda_n-\epsilon_n}\,(1+o(1)).$$

Consider $\epsilon_n = \sqrt{cn\ln n}$. Under the conditions of Lemma 4.2 on $\lambda_n$, $\epsilon_n = o(\lambda_n)$, and one has

$$\kappa_{n1} < e^{-\frac{c}{2}\cdot\frac{n\ln n}{n}(1+o(1))} = n^{-\frac{c}{2}(1+o(1))}.$$

Thus, for any $c>4$, $\sum_n n\kappa_{n1} < \infty$ and, therefore, $\sum_n S_{n1} < \infty$ since the factor $e^{1/(12(n-\lambda_n))}\sqrt{\frac{n-\lambda_n}{n-\lambda_n-\epsilon_n}}$ in (6.5) tends to one. Similarly, for $S_{n2}$, one has

$$\begin{aligned} S_{n2} &= \frac{\sum_{k=n-\lambda_n+\epsilon_n}^{n} b_k}{\sum_{k=0}^{n} b_k} < \frac{nb_{n-\lambda_n+\epsilon_n}}{b_{n-\lambda_n}} = \frac{n(n-\lambda_n)^{\epsilon_n}(n-\lambda_n)!}{(n-\lambda_n+\epsilon_n)!} \\ &< n\kappa_{n2}\sqrt{\frac{n-\lambda_n}{n-\lambda_n+\epsilon_n}}\, e^{1/(12(n-\lambda_n))}, \end{aligned}$$

where

$$\kappa_{n2} = \left(\left(1-\frac{\epsilon}{n-\lambda_n+\epsilon_n}\right)^{(n-\lambda_n+\epsilon_n)/\epsilon_n} e\right)^{\epsilon_n},$$

and, therefore,

$$\ln\kappa_{n2} = -\frac{1}{2}\,\frac{\epsilon_n^2}{n-\lambda_n+\epsilon_n}\,(1+o(1)).$$

Repeating the same arguments for $\kappa_{n2}$, for $\epsilon_n = \sqrt{cn\ln n}$, one has

$$\kappa_{n2} < n^{-\frac{c}{2}(1+o(1))},$$

and $\sum_n S_{n2}$ also converges for $c>4$. □

*Acknowledgment.* The authors would like to thank Yoav Benjamini and Vadim Grinshtein for fruitful discussions, and the two anonymous referees for their helpful comments.

## References


BARNDORFF-NIELSEN, O.E. and COX, D.R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London.

BAYARRI, M.J. and BERGER, J.O. (2004). The interplay of Bayesian and frequentist analysis. *Statist. Science*, **19**, 58–80.

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., KRIEGER, A.M. and YEKUTIELI, D. (2006). Adaptive linear step-up and step-down procedures that control the false discovery rate, *Biometrika* **93**, 491–507.

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency, *Ann. Statist.* **29**, 1165–1188.

BERGER, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer, New York.

BERGER, J.O. (2003). Could Fisher, Jeffreys and Neymann have agreed on testing? (with discussion), *Statist. Science* **18**, 1–32.

BERGER, J.O., BOUKAI, B. and WANG, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis, *Statist. Science* **12**, 133–160.

BERGER, J.O. and PERICCHI, L.R. (1996). The intrinsic Bayes factor for model selection and prediction, *J. Amer. Statist. Assoc.* **91**, 109–122.

BERRY, D.A. and HOCHBERG, Y. (1999). Bayesian perspectives on multiple comparisons, *J. Statist. Plann. Infer.* **82**, 215–227.

CASELLA, G. and BERGER, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem, *J. Amer. Statist. Assoc.* **82**, 106-111.

GOOD, I.J. (1992). The Bayes/non-Bayes compromise: A brief review, *J. Amer. Statist. Assoc.* **87**, 597–606.

GOPALAN, R. and BERRY, D.A. (1998). Bayesian multiple comparisons using Dirichlet process priors, *J. Amer. Statist. Assoc.* **93**, 1130–1139.

HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.

HOCHBERG, Y. and TAMHANE, A.C. (1987). *Multiple Comparisons Procedures*. Wiley, New York.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure, *Scand. J. Statist.* **6**, 65–70.

HSU, J.C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC, Washington D.C..

ISHWARAN, H. and RAO, J.S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection, *J. Amer. Statist. Assoc.* **98**, 438–455.

KASS, R.E. and RAFTERY, A.E. (1995). Bayes factors, *J. Amer. Statist. Assoc.* **430**, 773–795.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison (with discussion), *J. Roy. Statist. Soc. Ser. B*, **56**, 99–118.

REINER, A., YEKUTIELI, D. and BENJAMINI, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics* **19**, 368–375.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician, *Ann. Statist.* **12**, 1151–1172.

Sarkar, S.K. (2002). Some results on false discovery rate in stepwise multiple testing procedures, *Ann. Statist.* **30**, 239–257.

Sarkar, S.K. and Chen, J. (2004). A Bayesian stepwise multiple testing procedure, *Technical Report, Temple University.*

Scott, J.G. and Berger, J.O. (2006). An exploration of aspects of Bayesian multiple testing, *J. Statist. Plann. Infer.* **136**, 2144–2162.

Storey, J.D. and Tibshirani, R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays, *Technical Report* 2001-28, Dep. of Statistics, Stanford University.

Waller, R.A. and Duncan, D.B. (1969). A Bayes rule for the symmetric multiple comparison problem, *J. Amer. Statist. Assoc.* **64**, 1184–1503.

Felix Abramovich
Department of Statistics
and Operations Research
Tel Aviv University
Tel Aviv 69978, Israel
E-mail: felix@post.tau.ac.il

Claudia Angelini
Istituto per le Applicazioni
del Calcolo Mauro Picone
Consiglio Nazionale delle Ricerche
80131 Napoli, Italy
E-mail: c.angelini@iac.cnr.it

Paper received October 2005; revised May 2006.