On the Empirical Bayes Approach to the Problem of Multiple Testing

Małgorzata Bogdan[1, 2, *, †], Jayanta K. Ghosh[2], Aleksandra Ochman[1] and Surya T. Tokdar[3]

[1] *Institute of Mathematics and Computer Science, Wrocław University of Technology, Wrocław, Poland*
[2] *Department of Statistics, Purdue University, West Lafayette, IN, U.S.A.*
[3] *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, U.S.A.*

We discuss the Empirical Bayes approach to the problem of multiple testing and compare it with a very popular frequentist method of Benjamini and Hochberg aimed at controlling the false discovery rate. Our main focus is the 'sparse mixture' case, when only a small proportion of tested hypotheses is expected to be false. The specific parametric model we consider is motivated by the application to detecting genes responsible for quantitative traits, but it can be used in a variety of other applications. We define some Parametric Empirical Bayes procedures for multiple testing and compare them with the Benjamini and Hochberg method using computer simulations. We explain some similarities between these two approaches by placing them within the same framework of threshold tests with estimated critical values.

KEY WORDS: multiple testing; Empirical Bayes; false discovery rate

1. INTRODUCTION

Due to the recent development of technology, scientists and engineers can now collect and store massive data sets. To analyze such data sets, many statistical hypotheses are often tested at once. One of the most prominent examples of such multiple testing procedures is the analysis of genetic microarrays, where the expression of thousands of genes is simultaneously quantified and the hypotheses concerning the involvement of these genes in the investigated biological process are verified. The microarray analysis triggered the development of new statistical methods aimed at controlling some measures of error of multiple testing procedures (see, e.g. References[1–6]). In the present paper, we will consider another important example of multiple testing in statistical genetics, namely the problem of locating genes responsible for certain quantitative traits (quantitative trait loci (QTL)). Apart from widening the general biological knowledge, QTL mapping is often used to detect genes influencing economically important traits in

*Correspondence to: Małgorzata Bogdan, Institute of Mathematics and Computer Science, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland.
†E-mail: malgorzata.bogdan@pwr.wroc.pl

domesticated animals and industrial plants (see, e.g. References[7–9]). In humans, QTL mapping is often applied to locate genes responsible for the development of some medical conditions (see, e.g. Reference[10] as well as examples and references in Reference[11]).

In order to locate QTLs, geneticists use molecular markers. These are pieces of DNA that exhibit variation between individuals. Their characteristics (i.e. genotypes) can be determined experimentally. From the statistical point of view, marker genotypes can be treated as qualitative explanatory variables. If a QTL is located close to a given marker, we expect to see an association between the marker genotype and the trait value. In case when there are only two possible marker genotypes, this association can be measured using the standard t test. The number of markers used in typical genome scans for QTL usually reaches several hundreds, so the number of tests performed is usually substantially smaller than in the case of microarrays. However, similarly as in microarrays, usually only a very small proportion of tested markers corresponds to real QTL.

In the current paper, similarly as in our previous work[12], we compare the frequentist and Bayesian approach to the related multiple testing problem. Compared with Reference[12], we put more emphasis on interpretations and explanations of the similarities and differences between the two approaches.

In the frequentist multiple testing, the main emphasis is put on designing efficient methods for adjusting the significance level of individual tests so as the control of the familywise error rate (FWER) or the false discovery rate (FDR) is obtained. Bayesian multiple testing procedures that usually aim at minimizing the Bayesian risk based on a certain loss function, do not require a specific adjustment for multiple testing. Bayesian methods are very flexible and can be used in the situation when the cost of missing the true signal is comparable or even larger than the cost of the false discovery.

The full Bayesian approach, based on minimizing the posterior Bayes risk, usually requires implementation of some kind of Markov chain Monte Carlo (MCMC) algorithm and may be very much computationally involved. In our previous article[12], we investigated the performance of some parametric and non-parametric Bayes procedures for multiple testing and demonstrated that in the parametric setting full Bayes procedures can be closely approximated by simple and quick Empirical Bayes methods. We also demonstrated that the Bayes multiple testing procedures compare favorably to the popular Benjamini and Hochberg (BH) procedure.

According to our earlier research both BH and Empirical Bayes methods are superior to the well-known Bonferroni multiple testing procedure, which tests each component hypothesis independently of others. The reason for the superiority of the Bayes methods comes from modeling by the so-called mixture models. Learning about the common hyperparameters in the mixture model provides additional information on each component test. This provides superior tests as in the Stein estimation through parametric empirical Bayes (PEB) methods (see Reference[13]). In the present paper, we explore similarities of the performance of BH and PEB methods for multiple testing. In particular we note that, similar to the considered PEB procedures, the BH procedure is an example of a threshold test with estimated critical values, dependent on the overall distribution of all test statistics. Compared to our earlier work we also define and investigate a new Empirical Bayes procedure aimed at controlling the positive false discovery rate (pFDR).

The outline of the paper is as follows. In Section 2, we introduce and interpret our statistical model. In Section 3, we discuss different notions of error in multiple testing and compare Bayesian and frequentist approach to this problem. In Section 4, we define the procedures considered in our study and discuss the problem of non-identifiability of model parameters. Section 5 contains the results of the simulation study, Section 6 gives the final conclusions and the Appendix contains some technical details on the computations of the threshold for the BH procedure.

2. STATISTICAL MODEL

Consider the problem of QTL mapping. Let X_i be the t statistic measuring the association between the ith marker and the trait. In typical QTL mapping experiments, the sample size exceeds 200. Therefore, usually

we may safely assume that the test statistics X_i have normal distributions $N(\mu_i, \sigma^2)$. We also assume that X_i, $1 \leq i \leq m$, are independent.

In a fully classical setup, μ_i's are unknown constants, $\mu_i = 0$ corresponds to the null distribution and describes the situation when the marker is not in the QTL neighborhood. In Bayesian and Empirical Bayesian approach, we treat unknown μ_i's as random variables and model them using some probability distribution. In the context of QTL mapping it is often assumed that under the alternative hypothesis $\mu_i \sim N(0, \tau^2)$ (see, e.g. Reference[14]). This assumption is appropriate in the situation when the prior probabilities of a positive and a negative QTL effect are the same and the QTL effects are comparable with each other. Under this assumption, the non-null distribution of X_i is $N(0, \sigma^2 + \tau^2)$. Additionally, we define a random indicator variable γ_i, which is equal to 1 if X_i is generated by the non-null distribution (i.e. it represents the signal) or 0 in the other case. If $p = P(\gamma_i = 1)$ is the fraction of markers in the QTL's neighborhood, then the marginal distribution of X_i is the scale mixture of normals, namely,

$$X_i \sim (1 - p)N(0, \sigma^2) + pN(0, \sigma^2 + \tau^2) \tag{1}$$

For each i, we test whether X_i has a null or non-null distribution, i.e.

$$H_{0i} : \gamma_i = 0 \quad \text{vs} \quad H_{Ai} : \gamma_i = 1 \tag{2}$$

In practical applications, parameters p and τ are usually unknown. As far as σ is concerned we will consider two cases: when $\sigma = 1$ or when it is unknown. Large sample sizes used in typical genome scans allow one to assume that under the null hypothesis the distribution of the t statistic is approximately $N(0, 1)$. The case of unknown σ corresponds to the situation when the purpose of the study is to distinguish strong QTL from the background of many genes with very small effects, called polygenes. This situation often occurs in practice, since the majority of complex traits are influenced by a very large number of polygenes. In this case, σ^2 takes into account both the variability of the error term and the unknown variability of the distribution of polygenic effects and needs to be estimated (or integrated out).

In the mixture model (1), a special role is played by the threshold tests of the form: reject H_{0i} iff $|X_i| > c$, where c is the same threshold for all i. This c will typically depend on the (hyper) parameters p, σ^2, τ^2 of the mixture model, hence there is a need to estimate it. Typically, one would estimate p, σ^2 and τ^2 and choose a plug in estimate $c(\hat{p}, \hat{\sigma}^2, \hat{\tau}^2)$. We may think of this as an adaptive threshold test. PEB tests in this paper are directly of this form. Moreover, we observe that the BH procedure can also be considered as an adaptive threshold test.

Remark 1. One of the aims of QTL mapping is a precise location of 'strong' genes influencing a given trait. In this situation, the number of interesting genes might be very small, usually not exceeding 10. To distinguish such genes from the background noise their effects μ_i need to be comparable or larger than most of the noise terms. Note that the expected value of the maximum of m independent $N(0, \sigma^2)$ variables is approximately equal to $\sigma\sqrt{2 \log m}$. Therefore, in our simulations, reported in next sections, we use $\tau = \sigma\sqrt{2 \log m}$.

Remark 2. The assumption of independence of X_i's can be safely used when the investigated markers are distant from each other. In the case when markers are tightly spaced, test statistics at neighboring markers may be strongly correlated. Bayesian approach to multiple testing is flexible enough to take the dependency between the test statistics into account (see e.g. Reference[15]). Developing appropriate, more sophisticated methods of QTL data analysis is the topic of our current research.

3. BAYESIAN AND FREQUENTIST APPROACH TO MULTIPLE TESTING

Following the notation of Reference[1], Table I defines variables describing counts of possible outcomes of the multiple testing procedure.

Table I. Counts of possible outcomes of m hypothesis tests

	Accept null	Reject null	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

The classical measure of error of the multiple testing procedure is the familywise error rate, $FWER = P(V>0)$. One of the most popular multiple testing procedures providing the control of FWER at the level α is the Bonferroni correction, which suggests testing each of the investigated hypotheses at the significance level α/m. The control of FWER is also provided by a stepwise multiple testing procedure of Holm[16], which is uniformly more powerful than the Bonferroni correction.

FWER is a very stringent measure of error. In many practical situations, the costs of not detecting the true signal may be larger than the cost of the false positive. Therefore, scientists are interested in statistical procedures that allow for some false discoveries, as long as they consist only a small proportion of all discoveries. Seeger[17] elaborates on the idea of Eklund (unpublished seminar papers) and discusses a stepwise multiple testing procedure aimed at controlling the proportion of false discoveries among all discoveries. The same stepwise multiple testing procedure was later discovered by Simes[18] and has been proved to control FWER in a weak sense (when all hypothesis are true). The notion of proportion of false discoveries appeared again in the seminal paper of Sorić[19]. Following this paper, Benjamini and Hochberg[1] formally defined the false discovery rate as $FDR = E(V/R)$, where $V/R = 0$ if $R = 0$. Benjamini and Hochberg also proved that the multiple testing procedure of Seeger and Simes controls FDR at a desired level, for every combination of true and false hypotheses. The Benjamini and Hochberg article came at about the same time as microarrays and triggered a great interest in FDR controlling methods. As a result, Seeger's procedure is currently known as the Benjamini and Hochberg (BH) procedure.

Let $P_{(1)} \leq P_{(2)} \cdots \leq P_{(m)}$ be the ordered p-values of m tests. Let

$$k = \max\left\{ i : P_{(i)} \leq \frac{i\alpha}{m} \right\} \tag{3}$$

BH rejects all hypotheses for which the corresponding p-values are smaller than $P_{(k)}$.

Remark 3. In Reference[20], it is proved that under the classical setup and when test statistics are continuous and independent, FDR of BH is equal to

$$FDR = E_{\tilde{\mu}}[(V/R) I(R>0)] = \alpha m_0/m \tag{4}$$

where $I(R>0)$ is the indicator of the set $R>0$ and $\tilde{\mu} = (\mu_1, \ldots, \mu_m)$. Equation (4) shows that BH is more conservative than necessary to control FDR at the level α. Some modifications of BH, based on different methods of estimating m_0, are discussed in References[12,21].

Asymptotic results proved in Reference[22] show that for large values of m BH works like a procedure with a fixed threshold value. The threshold for $|X_i|$ depends on α and p as well as on the null and the alternative distributions of the test statistics according to the formula

$$c_{GW}(p, \sigma, \tau, \alpha) = inf\left\{ x : \frac{P_0(|X_i|>x)}{(1-p)P_0(|X_i|>x) + pP_A(|X|>x)} \leq \alpha \right\} \tag{5}$$

where P_0 and P_A are the probabilities corresponding to the null $N(0, \sigma^2)$ and alternative $N(0, \sigma^2 + \tau^2)$ distributions, respectively. As shown in Reference[3] BH is actually equivalent to using threshold (5), with the denominator estimated by $1 - F_m(x)$, where F_m is the empirical distribution function of $|X_i|$, $1 \leq i \leq m$.

Remark 4. Note that for very small p, the significant effects will usually be situated only in a tail of the empirical distribution. Since the relative error of the approximation of the cumulative distribution function by the empirical distribution function is large in the tails of the distribution, we expect that the asymptotic threshold (5) would not be accurate for very small p's. Our extensive simulations show that threshold (5) works well when $pm>10$. On the other hand, when $p=0$ the threshold of BH can be well approximated by the threshold of the Bonferroni correction. Our simulation study showed that in the intermediate range of pm, BH can be well approximated by the test procedure with the threshold value obtained by a linear interpolation between the asymptotic threshold (5) and the threshold of the Bonferroni correction (see the Appendix for more details on this approximation).

Remark 5. A striking result of Reference[23] ensures that the BH procedure is an adaptive minimax rule for many cases of sparse signals with small p, large m and $mp \to \infty$, the sort of situation described in this paper. However, in Reference[24] it is shown that BH fails to be optimal in the extremely sparse case, when $p \sim 1/m$, so mp does not tend to infinity. In general, i.e. for non-sparse problems, the decision theoretic implications of FDR controlling multiple testing procedures are unclear.

There is a good discussion of FDR and other similar measures in References[1,4]. In the case of mixture model, Storey[4] notes that several alternatives to FDR reduce to Storey's pFDR which is defined as $E(V/R|R>0)$. In Reference[4], it is shown that pFDR equals to BFDR (a sort of Bayesian FDR), defined as $BFDR = P(H_0 \text{ is true} \mid H_0 \text{ is rejected})$. For a threshold test with a critical value c this can be readily evaluated as

$$pFDR(c, p, \sigma, \tau) = \frac{(1-p)P_0(|X|>c)}{(1-p)P_0(|X|>c) + pP_A(|X|>c)} \tag{6}$$

In case of our testing problem (2) pFDR can be controlled at any given level α if only $p>0$. The critical values of the corresponding test $c_{pf}(p, \sigma, \tau, \alpha)$ can be computed numerically according to the formula

$$c_{pf}(p, \sigma, \tau, \alpha) = inf\{x : pFDR(x, p, \sigma, \tau) \leq \alpha\} \tag{7}$$

Figure 1(a) demonstrates critical values of the pFDR controlling rule when $\sigma = 1$, $\tau = \sqrt{2\log(200)}$ and $p \in [0, 0.2]$.

Remark 6. A comparison of (6) and (5) demonstrates that under the mixture model BH asymptotically controls pFDR at the level $(1-p)\alpha$. However, as illustrated by our simulation study, in case of very small p, pFDR of BH might be much larger than α.

Following Reference[12], we will now consider the multiple testing problem from the perspective of decision theory. Table II defines the specific matrix of losses for making the wrong decision.

Let us denote by t_1 and t_2 the probability of the type I and type II errors of a single test. The integrated Bayes risk related to the above matrix of losses is given by the following equation:

$$BR_{\delta_0,\delta_A} = \delta_0(1-p)t_1 + \delta_A p t_2 \tag{8}$$

For our model (1) the Bayes rule, i.e. the test that minimizes this risk, rejects the null hypothesis if

$$\frac{\phi_A(X_i)}{\phi_0(X_i)} > \frac{(1-p)\delta_0}{p\delta_A} \tag{9}$$

where ϕ_A and ϕ_0 are density functions of $N(0, \sigma^2+\tau^2)$ and $N(0, \sigma^2)$. Simplifying (9), the null hypothesis is rejected if

$$|X_i| > c_o(p, \sigma, \tau, \lambda) \tag{10}$$

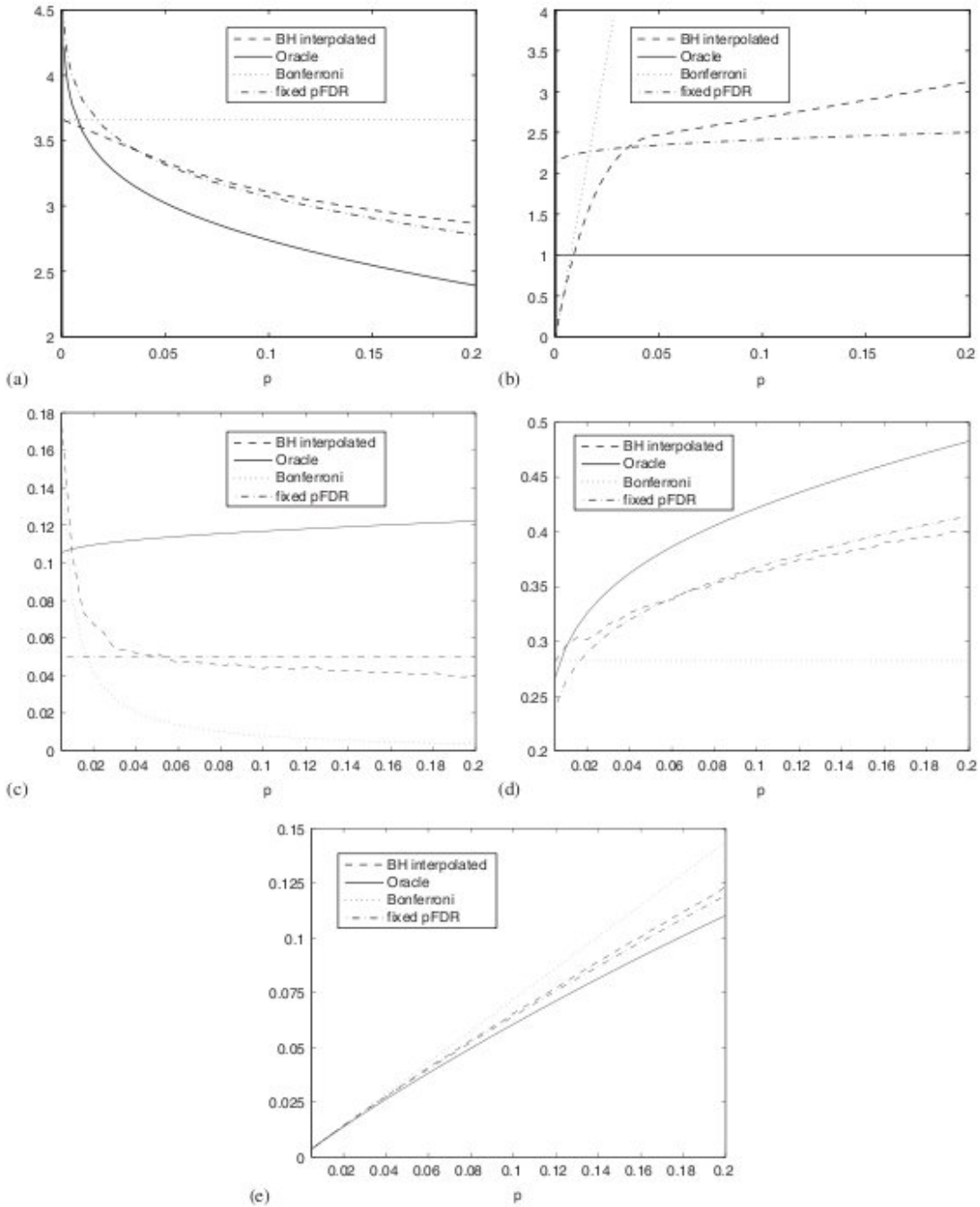

Figure 1. Characteristics of multiple testing procedures: (a) critical values; (b) loss ratio; (c) pFDR; (d) power; and (e) MP

Table II. Matrix of losses

	Accept H_{0i}	Reject H_{0i}
H_{0i} true	0	δ_0
H_{Ai} true	δ_A	0

where

$$c_o^2(p, \sigma, \tau, \lambda) = 2\frac{\sigma^2(\sigma^2+\tau^2)}{\tau^2}\log\left(\lambda\sqrt{1+\frac{\sigma^2}{\tau^2}}\frac{1-p}{p}\right) \tag{11}$$

with $\lambda = \delta_0/\delta_A$.

We refer to λ as the loss ratio. The bigger the value of λ the greater the importance given to type I error. When $\delta_0 = \delta_A = 1$ then the Bayes risk (8) is equal to the misclassification probability, *MP*. We call the corresponding Bayesian rule a *Bayes oracle* and compare other tests to this oracle.

Equation (11) allows to compute λ as a function of c, p, σ and τ for any test with a fixed threshold value c. Figure 1 contains graphs representing the critical values, loss ratios, pFDR, power and misclassification probabilities of all considered procedures. The purpose of this figure is to illustrate the potential of the Bayes oracle and pFDR controlling rule in the most advantageous situation, when all the parameters of the mixture distribution (1) are known.

Figure 1 demonstrates several interesting phenomena, which we briefly discuss. For the Bayes oracle we chose $\lambda = 1$, which corresponds to assigning the same importance to type I and type II errors. Interestingly, the loss ratio for pFDR controlling rule, reported in Figure 1(b), is very stable and in the range of $p \leq 0.2$ it takes values between 2.1 and 2.5. The loss ratio for BH varies from 0 for $p = 0$ to 3.2 for $p = 0.2$ and in the range of $p \in [0.05, 0.2]$ it is rather stable and can be well described as a linear function of p. The largest variability of λ is observed for the Bonferroni correction, which for large p assigns much larger weight to the type I error than to the type II error. Unless p is very small Bonferroni correction is also the most conservative and has the largest misclassification probability (*MP*).

Figure 1(e) shows that in the most interesting range of $p \leq 0.2$ MPs of Bayes oracle, BH and pFDR controlling rule are very similar. This suggests that the MP, which is a weighted sum of type I and type II errors, is not very sensitive to the choice of λ. This is, however, not true about pFDR, which depends on the ratio of type I and type II errors. The bigger value of $\lambda \approx 2.3$ for pFDR controlling rule leads to decrease in pFDR by half when compared with the oracle (see Figure 1(c)). As expected the pFDR of BH is substantially larger than 0.05 when $p < 0.02$. On the other hand, for $p \in [0.04, 0.2]$ BH is only slightly more conservative than pFDR controlling rule. Bayes oracle has the largest pFDR and the largest power. As expected, type I and type II errors balance in such a way that the *MP* of the Bayes oracle is smaller than of any other method.

4. EMPIRICAL BAYES PROCEDURES

The natural way of applying the Bayes oracle (10) or pFDR controlling rule (7) in the situation when parameters of (1) are unknown is to use some consistent estimates and plug them into (11) and (7). In particular, MLE estimators could be considered. However, the results of extensive computer simulations reported in Reference[12] show that the corresponding Empirical Bayes rules perform very poorly in the most interesting range of very small values of p. In Reference[12], it is argued that this poor behavior is the result of the problem with identifiability of model parameters. The mixture densities with p very close to 0 or 1 are very similar to the normal density and very difficult to distinguish using only the likelihood function.

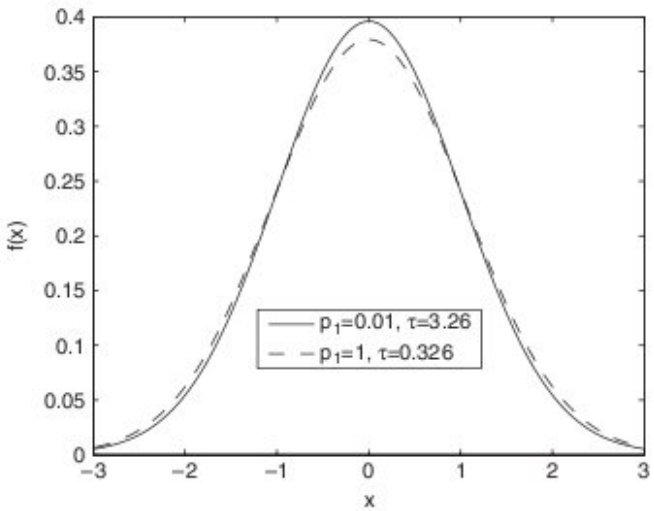

Figure 2. Similarity of different mixture models

An example of this phenomenon is illustrated in Figure 2, which shows that the density function of the mixture distribution (1) with $p = 0.01$, $\sigma = 1$ and $\tau = 3.26$ is very similar to the density with $p = 1$, $\sigma = 1$ and $\tau = 0.326$. Note that while these models seem to be equally plausible for describing the data they result in very different testing procedures. The second model implies that all the hypotheses are false and should be rejected, while the first model states that 99% of hypotheses are true and one should be very careful with rejections.

Bayesian statistics allows us to solve problems with identifiability of model parameters by utilizing the prior information. In Reference[12], a suitable method for estimation of model parameters was proposed. The method uses the prior knowledge on the expected number of QTL. In Reference[12] a corresponding PEB version of the oracle, PEB2, was proposed and investigated. We now review the estimation technique and define PEB2 as well as related versions of the pFDR controlling rule and the BH procedure.

Our method of estimation of model parameters uses the subjective, informative prior on p proposed in Reference[6], with the density

$$f(p) = \beta(1-p)^{\beta-1} \tag{12}$$

To adjust to the sparsity typical for QTL mapping experiments, we use $\beta = 22.76$. Then, the median of (12) is equal to 0.03, which for $m = 200$ corresponds to six signals on average.

Let

$$L(X_1, \ldots, X_m | p, \tau, \sigma) = \prod_{i=1}^{m} (p\phi_A(X_i) + (1-p)\phi_0(X_i))$$

where ϕ_A is a density of $N(0, \tau^2 + \sigma^2)$ and ϕ_0 is a density of $N(0, \sigma^2)$.

For each given p, we estimate $\tau(p)$ and $\sigma(p)$ using the second and the fourth moment of the mixture distribution. The resulting estimates are denoted by $\hat{\tau}(p)$ and $\hat{\sigma}(p)$. We observed that using the fourth moment makes our procedure sensitive to the change in the tail of the mixture distribution and in a very sparse mixture case gives better results than the maximum likelihood method.

Then, the estimate of p is obtained by maximizing

$$\log L(X_1, \ldots, X_m | p, \hat{\tau}(p), \hat{\sigma}(p)) - (\beta - 1)\log(1 - p) \tag{13}$$

and can be interpreted as a mode of the 'posterior' density of p. Let us denote this estimate by $\hat{p}$.

The Empirical Bayes version of Bayesian oracle, PEB2, is obtained by plugging $\lambda = 1$ and the estimates $\hat{p}$, $\hat{\tau}(\hat{p})$ and $\hat{\sigma}(\hat{p})$ into the formula for the critical value of the Bayes oracle (11). Similarly, the Empirical Bayes version of pFDR controlling rule, pFDR2, is obtained by plugging $\alpha = 0.05$ and the parameter estimates into the formula for the pFDR critical values (7). We also use $\hat{\sigma}(\hat{p})$ to compute p-values needed for the BH procedure.

When $\sigma = 1$ is known, the PEB2 and pFDR2 procedures are constructed accordingly. In that case $\tau(p)$ is estimated using the fourth moment of the mixture distribution.

Remark 4. In the full Bayes approach the nuisance parameters, like p, σ and τ, are integrated out with respect to a certain prior distribution. This typically requires an implementation of some numerical or stochastic methods to compute the related integrals. The full Bayes analysis of our testing problem was proposed and investigated in Reference[6]. This approach allows us to obtain posterior distributions of parameters of interest, μ_i, and so gives more information than our Empirical Bayes methods. However, the results of our simulations reported in Reference[12] show that in terms of the classifying decisions our relatively simple methods perform very similarly to the full Bayes approach.

5. RESULTS

To investigate the properties of our testing rules, we performed computer simulations. We considered the most interesting range of $p \leq 0.2$. For each p, we simulated 10 000 replicates of the random vector, consisting of $m = 200$ test statistics generated from the mixture distribution (1), with $\sigma = 1$ and $\tau = \sqrt{2\log(200)} \approx 3.26$. We used these replicates to estimate pFDR and the 'efficiency' of all testing procedures as well as to compute some characteristics of the distribution of the estimates of model parameters. The 'efficiency' of a testing procedure is defined with respect to the oracle (9) as *eff* = MP of the oracle/MP of a given procedure. The simulation results are presented in Figures 3 and 4.

Figures 3(a) and 4(a) show that the strong prior assumption on p allows for an accurate estimation of p in the most difficult cases of $p \leq 0.02$ and leads to some underestimation of this parameter for larger p. Interestingly, the bias of the estimate of p is smaller in case when σ is unknown. However, the standard deviation of the estimate of p and the corresponding mean square error in case when σ is unknown are substantially larger than when σ is known. The estimate of τ is almost unbiased in the entire range of $p \in [0.05, 0.2]$ (see Figures 3(b) and 4(b)). Some bias and a large standard deviation of the estimate of τ for $p < 0.05$ result from setting $\hat{\tau} = 0$ whenever $\hat{p} = 0$, which often occurs when the true p is close to 0. Due to a relatively good performance of the estimation technique PEB2 performs very well and in the range of $p \leq 0.2$ its efficiency is at the level close to 99% when σ is known (Figure 3(c)) and to 95% when σ is unknown (Figure 4(c)). pFDR seems to be more sensitive to the error of the estimation of p than the MP and rather difficult to control precisely. As shown in Figures 3(d) and 4(d), for $p = 0.005$ pFDR of pFDR2 is close to 0.09. This is, however, still much below the corresponding value for BH. When σ is known pFDR2 keeps pFDR at the level close to 0.04 for $0.02 < p \leq 0.2$. When σ is unknown pFDR2 is slightly more conservative and keeps pFDR at the level close to 0.03. As expected, for very small p ($p \leq 0.02$) pFDR of BH substantially exceeds 0.05. However when $p \in [0.04, 0.2]$ BH performs very similar to pFDR2 and, in case when σ is known, its pFDR is even closer to the nominal value of 0.05. Interestingly, also PEB2 has very good properties in terms of pFDR, which keeps it at the level close to 0.09 for all $0.03 < p < 0.2$.

It is interesting to observe that for $p \leq 0.03$ the efficiencies of pFDR2 and BH are slightly larger than the corresponding efficiency of PEB2. When $p \approx 0.005$ BH is almost optimal in terms of MP, with 'efficiency'

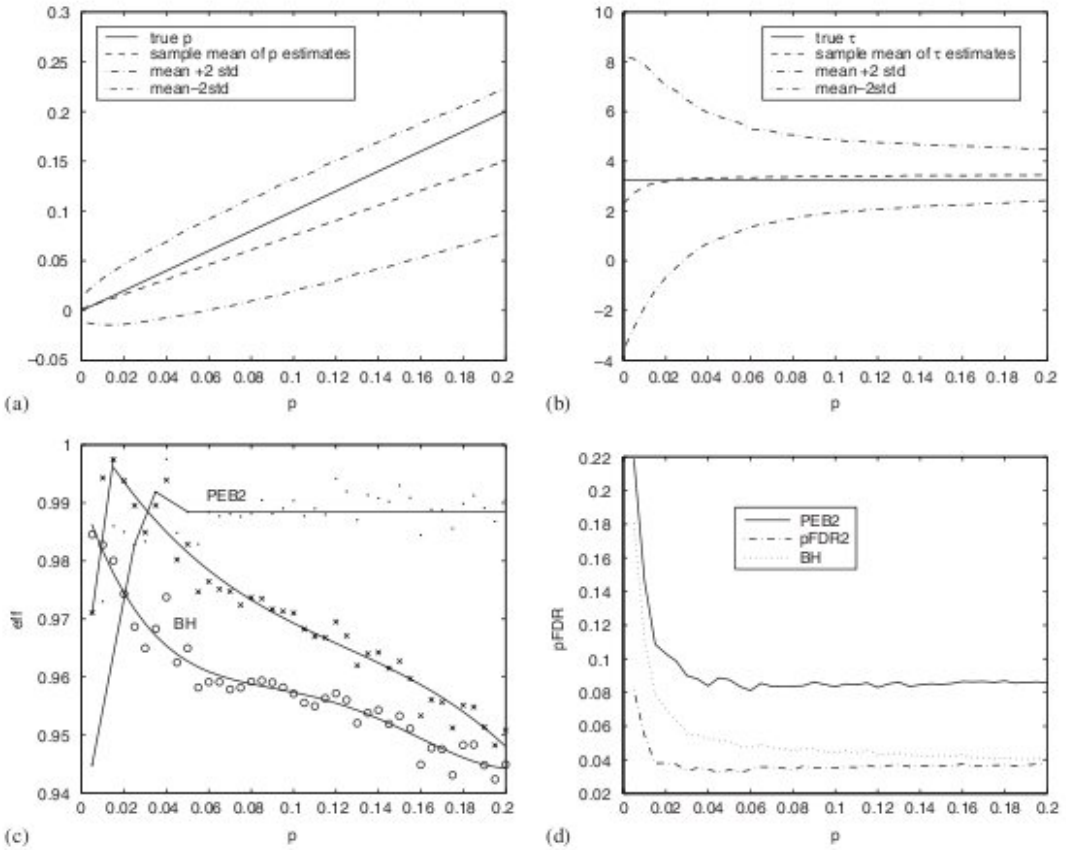

Figure 3. Characteristics of multiple testing procedures when σ is known: (a) distribution of estimates of p; (b) distribution of estimates of τ; (c) efficiency; and (d) pFDR

close to 1. However, this characteristic of BH systematically decreases with an increase of p, and at $p = 0.2$ it falls to 95% when σ is known and to 88% when σ is unknown. When $p \geq 0.03$, PEB2 performs systematically much better than BH and pFDR2 in terms of MP.

6. CONCLUSIONS

In this paper, we present the Empirical Bayes approach to multiple testing and compare some Empirical Bayes procedures to the well-known BH procedure. The main advantage of the presented Bayesian approach relies on the possibility of taking into account real costs of missing the true signal or making the false discovery. Bayesian methods also allow us to use the prior knowledge on the number of signals, which helps to solve the problem of non-identifiability of parameters of the mixture distribution. The results of our simulation study demonstrate good properties of the proposed methods. Particularly interesting is a very good performance of the Bayes classifier PEB2 both in terms of the MP and pFDR. Our simulations also demonstrate a very good behavior of the BH procedure, which in the considered range of p turns out to be very similar to the PEB procedure aimed at controlling pFDR. We explain this phenomenon by observing that the BH procedure is also an example of a threshold test with estimated critical values, dependent on the values of all test statistics. Our observations go along the discussion in Reference[25],

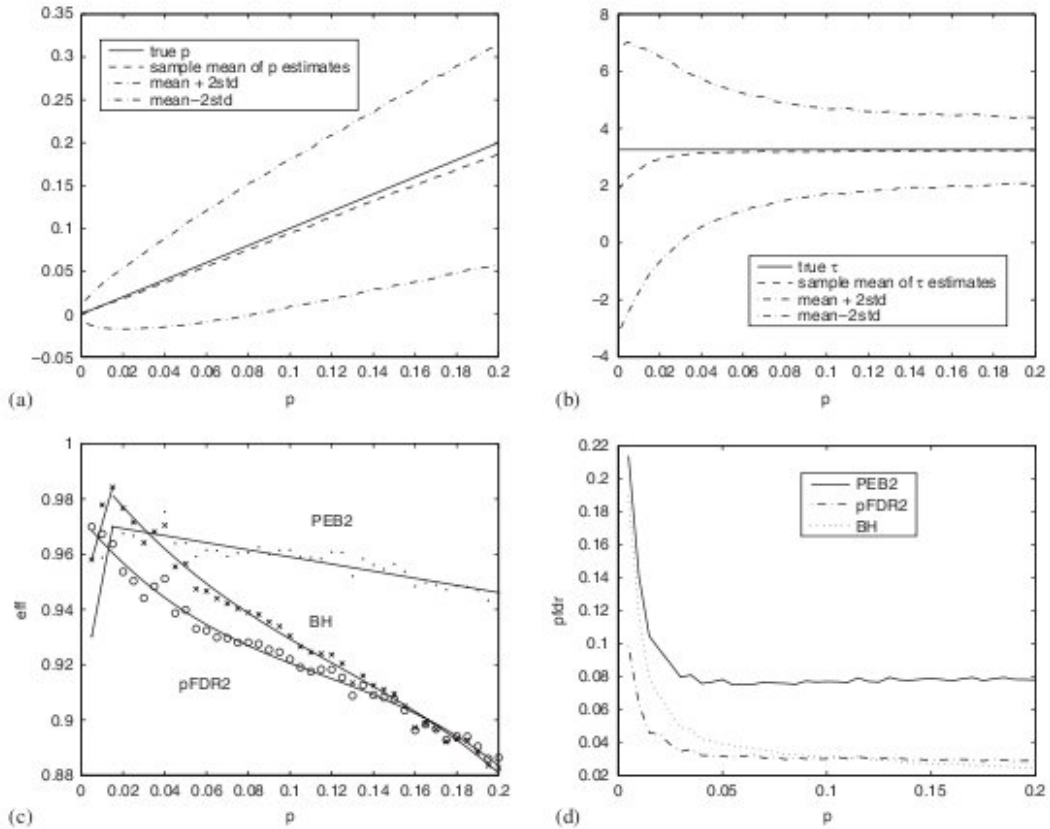

Figure 4. Characteristics of multiple testing procedures when σ is unknown: (a) distribution of estimates of p; (b) distribution of estimates of τ; (c) efficiency; and (d) pFDR

where some similarities between BH and the Empirical Bayes approach to multiple testing were pointed out. Another interesting phenomenon observed in the present paper is a relatively good performance of pFDR controlling rules in terms of MP. This can be explained by the stable ratios of losses for these procedures, which in the considered range of p are not much different than the loss ratio of a Bayes oracle.

Our simple Empirical Bayes methods can be seen as an approximation to the full Bayes approach, presented in Reference[6]. Their simplicity allows us to perform large-scale simulation studies. However, the full Bayes approach has the advantage of providing the posterior distributions for the parameters of interest, which allows for a more detailed analysis of the investigated problem.

The simulations presented in this paper were performed under the assumed model (1). Due to large sample sizes used in QTL mapping, the assumption of normality of X_i is often satisfied. According to our preliminary simulations, partially reported in Reference[12], in the sparse mixture case our PEB methods are robust to the deviations from the assumption of normality of μ_i. However, a detailed analysis of this problem still needs to be performed. In our simulation study, we did not consider the situation when the statistics X_i are correlated. We believe that weak violations of independence will not lead to strong influence on the outcome of our procedures. The detailed analysis of this problem as well as the development of PEB methods that would take into account the correlation structure typical for QTL mapping experiments with densely spaced markers is the topic of our current research.

Acknowledgements

We thank two anonymous referees for helpful comments and suggestions. The research of Małgorzata Bogdan was supported by grant 1 P03A 01430 of the Polish Ministry of Science and Higher Education.

REFERENCES

1. Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 1995; **57**:289–300.
2. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001; **96**:1151–1160.
3. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* 2002; **23**:70–86.
4. Storey JD. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* 2003; **31**:2013–2035.
5. Müller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association* 2004; **99**:990–1001.
6. Scott JG, Berger JO. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference* 2006; **136**:2144–2162.
7. Chardon F, Virlon B, Moreau L, Falque M, Joets J, Decousset L, Murigneux A, Charcosset A. Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics* 2004; **168**:2169–2185.
8. Khatkar MS, Thomson PC, Tammen I, Raadsma HW. Quantitative trait loci mapping in dairy cattle: Review and meta-analysis. *Genetic Selection Evolution* 2004; **36**:163–190.
9. Walling GA, Visscher PM, Andersson L, Rotschild MF, Wang L, Moser G, Groenen MAM, Bidanel J-P, Cepica S, Archibald AL, Geldermann H, de Koning DJ, Milan D, Haley CS. Combining analyses of data from quantitative trait loci mapping studies: Chromosome 4 effects of porcine growth and fatness. *Genetics* 2000; **155**:1369–1378.
10. Dick DM, Foroud T. Genetic strategies to detect genes involved in alcoholism and alcohol-related traits. *Alcohol Research and Health* 2002; **26**:172–180.
11. Lynch M, Walsh B. *Genetics and Analysis of Quantitative Traits*. Sinauer: Sunderland, MA, 1998.
12. Bogdan M, Ghosh JK, Tokdar ST. A comparison of the Simes–Benjamini–Hochberg procedure with some Bayesian rules for multiple testing. *Institute of Mathematical Statistics volume for Prof. P. K. Sen* 2007, to appear.
13. Efron B, Morris C. Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association* 1975; **68**:117–130.
14. Yi N. A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* 2004; **167**:967–975.
15. Broët P, Richardson S. Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics* 2006; **22**:911–918.
16. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
17. Seeger P. A note on a method for the analysis of significance en masse. *Technometrics* 1968; **10**:586–593.
18. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**:751–754.
19. Sorić B. Statistical 'discoveries' and effect-size estimation. *Journal of the American Statistical Association* 1989; **84**:608–610.
20. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 2001; **29**:1165–1188.
21. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery fate in multiple testing with independent statistics. *Journal of Educational Behavioral Statistics* 2000; **25**:60–83.
22. Genovese C, Wasserman L. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B* 2002; **64**:499–517.
23. Abramovich F, Benjamini Y, Donoho DL, Johnstone IM. Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics* 2006; **34**:584–653.
24. Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* 2004; **32**:962–994.
25. Efron B. Robbins, Empirical Bayes and microarrays. *Annals of Statistics* 2003; **31**:366–378.

APPENDIX. INTERPOLATED CRITICAL VALUES FOR BH PROCEDURE

For $\sigma = 1$, $m = 200$ and $\alpha = 0.05$ the threshold of the Bonferroni procedure is equal to

$$c_B = \Phi^{-1}\left(1 - \frac{0.05}{400}\right) \approx 3.6623$$

where Φ is the cdf of $N(0, 1)$.

According to our simulations, for $m = 200$ the asymptotic critical value for BH provided in (5) works well when $p \geq 0.05$. However, it is easy to see that when $p \to 0$ than $c_{GW} \to \infty$. This contradicts the fact that BH is always more liberal than Bonferroni correction. According to our simulations, when $p \in [0, 0.05]$, BH can be well approximated by the threshold test with the critical value obtained by a linear interpolation between c_B and $c_{GW}(0.05, \sigma, \tau, \alpha)$.

For $\sigma = 1$, $\tau = \sqrt{2\log(200)}$ and $\alpha = 0.05$, $c_{GW}(0.05, \sigma, \tau, \alpha) \approx 3.3326$. Thus for values of $p \in [0, 0.05]$, we approximate the threshold of BH by

$$c_{BH} = \frac{3.6623(0.05 - p) + 3.3326p}{0.05}$$

The graph of these critical values is provided in Figure 1(a).

Authors' biographies

Małgorzata Bogdan is an Assistant Professor in the Institute of Mathematics and Computer Science of Wrocław University of Technology in Poland. In 2000, she spent 8 months as a Visiting Scholar in the Department of Statistics of the University of Washington. She also often visits and teaches for the Department of Statistics at Purdue University. Her research interests include model selection criteria and multiple testing in high dimensions, Bayesian statistics, asymptotics, goodness-of-fit testing, statistical genetics and bioinformatics. She is a member of the Institute of Mathematical Statistics, the International Society for Bayesian Analysis and the European Network for Business and Industrial Statistics.

Jayanta K. Ghosh is a Professor in the Department of Statistics at Purdue University. He had been the Director and Jawaharlal Nehru Professor at the Indian Statistical Institute and President of the International Statistical Institute. He has been the editor of Sankhya and served on the editorial boards of several journals including the Annals of Statistics. Apart from Bayesian analysis, his interests include asymptotics, stochastic modeling, high-dimensional model selection, reliability and survival analysis and bioinformatics.

Aleksandra Ochman is a PhD student in the Institute of Mathematics and Computer Science of Wrocław University of Technology in Poland. The topic of her research is the Bayesian approach to the problem of multiple testing.

Surya T. Tokdar is a Morris H. DeGroot Visiting Assistant Professor in the Department of Statistics, Carnegie Mellon University. He graduated in 2006 from the Department of Statistics, Purdue University. He is the author or coauthor of four articles in distinguished statistical journals. His area of interest is Bayesian non-parametric and semiparametric analysis in the context of density estimation and regression.