Tail posterior probability for inference in pairwise and multiclass gene expression data

N. Bochkina, S. Richardson

Centre for Biostatistics, Imperial College, London W2 1PG

Summary. We consider the problem of identifying differentially expressed genes in microarray data in a Bayesian framework with a non-informative prior distribution on the parameter quantifying differential expression. We introduce a new rule, *tail posterior probability*, based on the posterior distribution of the standardised difference, to identify genes differentially expressed between two conditions, and we derive a frequentist estimator of the false discovery rate associated with this rule. We compare it to other Bayesian rules in the considered settings. We show how the tail posterior probability can be extended to testing a compound null hypothesis against a class of specific alternatives in multiclass data.

Key words: Bayesian analysis; tail posterior probability; compound hypothesis; differential expression; equivalence of Bayesian and frequentist inference; microarray gene expression; multiclass data.

1. Introduction

Studying the expression of genes simultaneously measured across several samples on gene expression microarrays has a wide range of applications including

email: n.bochkina@imperial.ac.uk

investigating genes interaction and changes of gene expression under different conditions or specific diseases. A general description of microarray experiments and associated statistical problems can be found, for instance, in McLachlan, Do, & Ambroise (2004).

Here we focus on the problem of differential expression which is a key question in many microarray studies. We consider it in a Bayesian framework which allows necessary flexibility for modelling diverse sources of variability in microarray data. Differential expression can be formulated as a hypothesis testing problem, which in Bayesian settings is often addressed by using a mixture prior for modelling the null and the alternative distributions of the difference δ_g between the log expressions of gene g in the two conditions, see, for example, Smyth (2004) and Gottardo et al. (2006). Under a mixture prior, decision rules based on Bayes factors (or empirical Bayes approximations of those) can be used for choosing differentially expressed genes but the mixture setup necessitates to choose how to model the alternative prior distribution. Baldi & Long (2001) model the difference using a Gaussian prior instead of the mixture prior and propose regularised t tests. Lewin et al. (2006) avoided the problem of specifying *a priori* a parametric form for the alternative distribution by using a non-informative prior for the difference δ_g and considering an interval null hypothesis of the form $|\delta_g| \leq \log_2 c$ based on interpreting δ_g as the log fold change and *a priori* information about the fold change of interest being above a set value.

However, in some cases, small but consistent fold changes are of interest and it is delicate to choose *a priori* a set value above which the expression changes are of interest. Motivated by the objective Bayesian approach pioneered by Jeffreys (1961), in this paper, we propose to use non-informative prior settings for δ_g and a data-related threshold which takes into account the variability of each gene leading to a new type of a selection rule which we refer to as *tail posterior probability*. We advocate that this selection rule, based on the standardised difference, is particularly appropriate in application to microarray data.

We also extend the tail posterior probability to multiclass data and show how it can address the problem of testing a compound null hypothesis, *i.e.* a hypothesis involving several comparisons of interest, against *structured alternative hypotheses*. This is an important property since it allows to evaluate directly the uncertainty associated with a biological question which in multiclass settings often involves several comparisons of interest.

The paper is organised as follows. In Section 2 we set out our Bayesian formulation. In Section 3, we discuss different types of possible selection rules, introduce the *tail posterior probability* and compare it to other methods for identifying differentially expressed genes in the considered Bayesian framework. In Section 4 we propose a frequentist estimator for the false discovery rate based on the tail posterior probabilities. In Section 5, we apply the proposed methodology to the microarray data. In Section 6 we extend the introduced tail posterior probability to test compound hypotheses in multiclass experiments and conclude with a brief discussion.

2. Bayesian framework

2.1 *Testing hypothesis of no differential expression*

We assume that intensity on the $\log_2$ scale y_{gsr} for gene $g = 1, \ldots, n$, condition $s = 1, 2$ and replicate $r = 1, \ldots, m_s$ is observed with normal errors:

$$y_{g1r} \sim N(\alpha_g - \delta_g/2, \sigma^2_{g1}), \quad y_{g2r} \sim N(\alpha_g + \delta_g/2, \sigma^2_{g2}). \tag{1}$$

We are interested in testing the null hypothesis related to the parameter of differential expression δ_g for each gene g:

$$H_0^{(g)}: \quad \delta_g = 0 \quad \text{versus} \quad H_1^{(g)}: \quad \delta_g \neq 0. \tag{2}$$

Now we specify the hierarchical model for the parameters of (1). Since in most cases we do not have any prior information about the distribution of the non-zero differences δ_g, we avoid using a mixture prior for δ_g, and consider a non-informative prior for both mean parameters α_g, δ_g, and the exchangeable distributions for the variances σ^2_{gs}, $s = 1, 2$:

$$\alpha_g, \, \delta_g \sim 1, \quad \sigma^2_{gs} \sim f(\sigma^2_{gs} \mid \beta_s), \quad \beta_s \sim f(\beta_s), \tag{3}$$

where β_s are the hyperparameters of the distribution of variance, with prior distributions $f(\beta_s)$ at the final level of the hierarchy. This hierarchical setup that embeds the exchangeability assumption, is commonly used to borrow information across variables or observations (see Chapter 5 in Gelman et al., 2004) and is needed here to stabilise variance estimates based on typically few replicates.

Although the priors for α_g and δ_g are improper, it can be easily shown that under model (1) with at least one observation in each condition, the resulting posterior distributions for α_g and δ_g are proper.

2.2 *Posterior distributions*

Since the model is gaussian, the likelihood depends on the following data summaries: sample means $\bar{y}_{gs} = \frac{1}{m_s}\sum_{r=1}^{m_s} y_{gsr}$ and sample variances $s_{gs}^2 = \frac{1}{m_s-1}\sum_{r=1}^{m_s}(y_{gsr} - \bar{y}_{gs})^2$, $s = 1, 2$.

Therefore, under the model described in Section 2.1, the sampling model for the sufficient statistics $\bar{y}_g = \bar{y}_{g2} - \bar{y}_{g1}$ and s_{gs}^2 for parameters δ_g and σ_{gs}^2 is given by

$$\bar{y}_g \mid \delta_g, w_g \sim N(\delta_g, w_g^2), \quad s_{gs}^2 \mid \sigma_{gs}^2 \sim \Gamma\left(\frac{\widetilde{m}_s}{2}, \frac{\widetilde{m}_s}{2\sigma_{gs}^2}\right), \tag{4}$$

resulting in the following posterior probability model:

$$\delta_g \mid \bar{y}_g, w_g \sim N(\bar{y}_g, w_g^2),\ \ f(\sigma_{gs}^2 \mid s_{gs}^2, \beta_s) \propto f_\gamma\left(s_{gs}^2 \mid \frac{\widetilde{m}_s}{2}, \frac{\widetilde{m}_s}{2\sigma_{gs}^2}\right) f(\sigma_{gs}^2|\beta_s), \tag{5}$$

where $\widetilde{m}_s = m_s - 1$, $w_g^2 = \sigma_{g1}^2/m_1 + \sigma_{g2}^2/m_2$ is the variance of $\bar{y}_g$ and $f_\gamma(x \mid a, b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$, $x > 0$, is the density of the gamma distribution $\Gamma(a, b)$. We condition on the hyperparameters β_s to avoid lengthy formulae; the fully integrated posterior distribution of σ_{gs}^2 conditioned on data can be easily obtained from (5). Note that the posterior distribution of w_g^2 depends on the data only through s_{gs}^2 and its density function $f(w_g^2 \mid s_{gs}^2)$ can be written in terms of the posterior distributions of σ_{gs}^2.

To obtain the posterior density functions of the difference δ_g and of the standardised difference $t_g = \delta_g/w_g$: $t_g \mid \bar{y}_g, w_g \sim N(\bar{y}_g/w_g, 1)$, we integrate the full conditional distribution of δ_g (5) and of t_g over w_g^2:

$$f(\delta_g \mid \bar{y}_g, s_{gs}^2) = \int_0^\infty \frac{1}{w_g}\varphi((\delta_g - \bar{y}_g)w_g^{-1}) f(w_g^2|s_{gs}^2) d(w_g^2), \tag{6}$$

$$f(t_g|\bar{y}_g, s_{gs}^2) = \int_0^\infty \varphi(t_g - w_g^{-1}\bar{y}_g) f(w_g^2|s_{gs}^2) d(w_g^2), \tag{7}$$

where $\varphi(x)$ (and $\Phi(x)$) is the density (and the cumulative distribution) function of the standard normal distribution. If we introduce the function $G(x,y) = \int_0^\infty \Phi(x + w_g^{-1}y) f(w_g^2|s_g^2) d(w_g^2)$, the cumulative posterior distribution functions of δ_g and t_g can be written as $P\{\delta_g < x \mid y_{gsr}\} = G(0, x - \bar{y}_g)$ and $P\{t_g < x \mid y_{gsr}\} = G(x, -\bar{y}_g)$ respectively.

3. Bayesian selection rules for pairwise comparisons

In this section we discuss and compare different rules for selecting differentially expressed genes. Each selection rule is associated with two choices: what parameter to use (e.g. δ_g or t_g) and what posterior probability event to consider, e.g. whether to use fixed *a priori* thresholds or data dependent ones. We will also compare the considered selection rules on the microarray data.

3.1 *Tail posterior probabilities*

To select differentially expressed genes, we propose to use selection rules of the form: $p(T_g, \theta) = P\{|T_g| > \theta \mid y_{gsr}\} \geq p_{cut}$, where T_g is a function of parameters, and we propose to use a data-dependent threshold $\theta = T_g^{(\alpha)}$ which is $1 - \alpha/2$ percentile of T_g under the (hypothetical) condition $\bar{y}_g = 0$. Such threshold is related to the variability of gene g and is preferable in cases where the cutoff on the fold change is not known *a priori*. In Section 3.2 we will introduce a loss function under which the rules of this type are optimal.

We start with the case $T_g = \delta_g$. To define the percentile, $\delta_g^{(\alpha)}$, we use the following heuristics. Suppose that we could have observed comparable data *with the same variability under the null hypothesis*, and, in ideal circumstances, that we would have observed, on average, a zero value for the sample mean difference $\bar{y}_g$. In this case, the posterior distribution $f(\delta_g|\bar{y}_g = 0, s_{gs}^2)$

could be used to define the percentile $\delta_g^{(\alpha)}$ of order $1-\alpha/2$ for (hypothetical) data with no differential expression. Given $\delta_g^{(\alpha)}$, we then define the *tail posterior probability* by:

$$p(\delta_g, \delta_g^{(\alpha)}) = P\{|\delta_g| > \delta_g^{(\alpha)} | y_{gsr}\}.$$

To find the percentiles $\delta_g^{(\alpha)}$ in general (*i.e.* for an arbitrary prior distribution of w_g^2), we need to calculate the distribution function of δ_g given s_{gs}^2 and $\bar{y}_g = 0$, $G(0, \delta_g)$, which involves numerical integration over the posterior distribution of w_g^2 for each gene and thus is computationally intensive.

If instead of δ_g, we consider T_g to be the standardised difference, $t_g = \delta_g/w_g$, we can obtain a straightforward expression for $t_g^{(\alpha)}$. Indeed, using (7) and a similar heuristic argument, we can see that given $\bar{y}_g = 0$ (*i.e.* for the average data under the null hypothesis), $f(t_g|\bar{y}_g = 0, w_g) = \varphi(t_g)$, *i.e.* that this distribution, and thus the distribution $f(t_g|\bar{y}_g = 0, s_{gs}^2)$, does not involve gene-specific parameters or data. Therefore the corresponding percentile $t_g^{(\alpha)}$ is easy to calculate: $t_g^{(\alpha)} = t^{(\alpha)} = \Phi^{-1}(1-\alpha/2)$.

We investigated performance of the tail posterior probabilities based either on δ_g or on t_g, and found them similar (results not shown). The advantage of using t_g is that the dependence of the corresponding threshold $\theta = w_g t^{(\alpha)}$ for δ_g, $\{|t_g| > t^{(\alpha)}\} = \{|\delta_g| > w_g t^{(\alpha)}\}$, on w_g is explicit, as opposed to the threshold $\delta_g^{(\alpha)}$ in the event $\{|\delta_g| > \delta_g^{(\alpha)}\}$ (see Section 3.2), which also makes the computation of the posterior probability of the latter event more difficult than in the former case. Therefore, throughout the paper we will concentrate on the tail posterior probability based on the standardised difference t_g:

$$p(t_g, t^{(\alpha)}) = P\{|t_g| > t^{(\alpha)} | y_{gsr}\}. \tag{8}$$

Note also that both tail posterior probabilities can be written in terms of function $G(x,y)$ defined in Section 2.2:

$$\begin{aligned} p(\delta_g, \delta_g^{(\alpha)}) &= G(0, -\delta_g^{(\alpha)} + \bar{y}_g) + G(0, -\delta_g^{(\alpha)} - \bar{y}_g), \\ p(t_g, t^{(\alpha)}) &= G(-t^{(\alpha)}, \bar{y}_g) + G(-t^{(\alpha)}, -\bar{y}_g). \end{aligned}$$

3.2 *Link between the tail posterior probabilities and decision theory*

Suppose that d_g is a decision whether to declare gene g to be differentially expressed ($d_g = 1$) or not ($d_g = 0$), and that we have the following loss function with a positive constant c:

$$L(\delta, \sigma^2, d) = c \sum_{g=1}^{n} d_g I\{|\delta_g| \le \theta(w_g)\} + \sum_{g=1}^{n} (1 - d_g) I\{|\delta_g| > \theta(w_g)\},$$

where $w_g^2 = \sigma_{g1}^2/m_1 + \sigma_{g2}^2/m_2$ as defined in Section 2.2, $I\{A\}$ is the indicator function of set A and $\theta(w_g)$ is a threshold which may depend on w_g. Then, following the approach of Müller et al (2004), it is easy to show that the optimal decision rule corresponding to the posterior expected loss function $EL(\delta, \sigma^2, d \mid y_{gsr})$ is $d_g = I\{P(|\delta_g| > \theta(w_g)|y_{gsr}) > c/(c+1)\}$. For example, for threshold $\theta(w_g) = \delta_g^{(\alpha)}$ which is independent of w_g, the optimal decision rule is expressed in terms of the tail posterior probability based on δ_g: $d_g = I\{p(\delta_g, \delta_g^{(\alpha)}) > c/(c+1)\}$, and for the threshold proportional to w_g, $\theta(w_g) = w_g t^{(\alpha)}$, the optimal decision is expressed in terms of the tail posterior probability based on t_g: $d_g = I\{p(t_g, t^{(\alpha)}) > c/(c+1)\}$. Thus, a similar structure of the loss function and the use of the quantile as a threshold imply close correspondence between the two considered tail posterior probability rules.

Tail posterior probabilities can also be viewed as a posterior discrepancy measure for model rejection which is valid in objective settings (Bernardo and Smith, 2004, Section 6.2.2), with discrepancy being $I\{|\delta_g| > \theta(w_g)\}$, with the respective choice of θ for each type of the tail posterior probability. Note that this discrepancy is similar to that for an interval null hypothesis, with the interval depending on the variability of δ_g. Bernardo and Smith also interpret the posterior discrepancy as a test statistic, and suggest to make a decision using its distribution under H_0 (in the frequentist sense). We follow this approach to derive a frequentist estimate of the false discovery rate associated with the tail posterior probability $p(t_g, t^{(\alpha)})$ in Section 4.2.

Interpreting a tail posterior probability as a test statistic, we can calibrate its distribution under H_0 in the frequentist sense. We find that

$$P\{p(\delta_g, \delta_g^{(\alpha)}) > 0.5 \mid H_0\} \approx \alpha, \tag{9}$$

i.e., that the quantile of order α of the distribution of $p(\delta_g, \delta_g^{(\alpha)})$ under H_0 is approximately 0.5 (see Web Appendix C). Since values of $p(\delta_g, \delta_g^{(\alpha)})$ and $p(t_g, t^{(\alpha)})$ are approximately the same, this property also holds for $p(t_g, t^{(\alpha)})$.

3.3 *A "p-value"-type selection rule*

In this section, we consider another Bayesian rule to select differentially expressed genes, the posterior probability that simply compares the parameter of interest δ_g to its value under the null hypothesis, *i.e.* compares δ_g to 0. We then show that it is equivalent to a frequentist test based on the marginal distribution of $\bar{y}_g$ in the considered framework.

If we are interested in the one-sided alternative $\delta_g > 0$, we can estimate its posterior probability $p(\delta_g, 0) = P\{\delta_g > 0 | y_{gsr}\}$, and for the complementary alternative, we can consider the posterior probability $P\{\delta_g < 0 | y_{gsr}\} = 1 - p(\delta_g, 0)$. Since the posterior probability of the two-sided alternative $P\{\delta_g \neq 0 | y_{gsr}\}$ under the considered hierarchical Bayesian model contains no information, to test the null hypothesis against this alternative, we take the maximum of the posterior probabilities of the one-sided alternatives: $\max\{p(\delta_g, 0), 1 - p(\delta_g, 0)\}$. The theorem below gives some theoretical results related to the distribution of these posterior probabilities under the null hypothesis and link these rules to frequentist testing procedures.

THEOREM 1. *In the settings considered above with non-informative priors for α_g and δ_g and an arbitrary hierarchical prior distribution for σ^2_{gs}, the following statements hold for any g.*
1. The distributions of $p(\delta_g, 0)$, $1-p(\delta_g, 0)$ and $2\max\{p(\delta_g, 0), 1-p(\delta_g, 0)\}-1$ are uniform on $[0, 1]$ under the null hypothesis.
2. Testing the null hypothesis $\delta_g = 0$ against alternatives $\delta_g > 0$, $\delta_g < 0$ or $\delta_g \neq 0$ using the posterior probabilities above is equivalent to testing the same hypotheses using corresponding p-values based on the marginal distribution of $\bar{y}_g$ under the null hypothesis.

Proof of the theorem generalised to the case where the distribution of $\delta_g - \bar{y}_g$, conditioned on data and the remaining parameters, is independent of $\bar{y}_g$, is given in the Appendix. To give an insight into the proof, we can rewrite the posterior probability $p(\delta_g, 0)$ in terms of function $G(x, y)$ defined in Section 2.2: $p(\delta_g, 0) = G(0, \bar{y}_g)$, and note that the cumulative distribution function of $\bar{y}_g$ given $\delta_g = 0$ is $P\{\bar{y}_g < x \mid \delta_g = 0, s^2_{gs}\} = G(0, x)$, as follows from (4), so the theorem becomes clear.

For example, in the conjugate case $\sigma^{-2}_{gs} = \sigma^{-2}_g \sim \Gamma(a, b)$, Smyth (2004) introduced the moderated t statistic $t^{mod}_g = \frac{\bar{y}_g\sqrt{k}}{S_g}$ for the frequentist testing of hypothesis (2), where $S^2_g = \frac{2b+s^2_g(m_1+m_2-2)}{2a+m_1+m_2-2}$ is the empirical Bayesian posterior estimate of the variance, $k = (m^{-1}_1 + m^{-1}_2)^{-1}$ and s^2_g is the pooled sample estimate of the variance. For fixed a, and b, t_g^{mod} has t distribution with $\nu = 2a + m_1 + m_2 - 2$ degrees of freedom under H_0, and it can be shown that the right-sided posterior probability is directly related to the corresponding p-value: $P\left\{\delta_g > 0 \mid \bar{y}_g, s_g^2, a, b\right\} = F_{T(\nu)}\left(\frac{\bar{y}_g\sqrt{k}}{S_g}\right) = 1 - P\left\{t > t_g^{mod}|H_0\right\}$.

If we consider an inverse gamma prior model for σ_g^2 with $a = b = 0$ (i.e. a non-hierarchical model with non-informative prior for the variances), the posterior probability $p(\delta_g, 0)$ will correspond to the p-value based on the standard t test. In the case of the small number of replicates m_s (e.g. for $m_1 = m_2 = 3$, as we have for the considered data), the gene-specific variance estimate will be s_g^2 which is unstable, and will lead to a high number of false discoveries (Baldi and Long, 2001).

3.4 *Comparison of selection rules*

In this section, we compare the gene lists obtained by using the tail posterior probability $p(t_g, t^{(\alpha)})$ based on the standardised difference, to gene lists based on other posterior probabilities, namely the posterior probability $p(\delta_g, \log_2 2)$ that the difference δ_g exceeds in absolute value an *a priori* specified threshold $\log_2 2$ (Lewin et al., 2006), and the posterior probability $p(\delta_g, 0)$ that the difference is greater than its null value. These posterior probabilities are calculated for one of the comparisons investigated in the experiment described in Section 5, namely for the comparison between control and insulin-treated H2Kb muscle cells in mouse at 2 hours, using 3 biological replicates in each condition. Hierarchical model defined by (1) and (3) with the inverse gamma prior distribution for equal variances $\sigma_{gs}^2 = \sigma_g^2$ was fitted using WinBUGS (see Section 5 for details of the hyperprior specification).

[Figure 1 about here.]

The criterion with the fixed threshold divides genes in three three biologically interpretable groups (see Figure 1c): genes whose 'true' fold change is greater than the threshold $\log_2 2$ with posterior probability close to 1 (differentially expressed), a large group of genes with 'true' fold change greater than $\log_2 2$ with posterior probability close to 0 (non-differentially expressed), and a small group of genes (compared to the other approaches) where there is substantial uncertainty whether their 'true' fold change is above or below the threshold (separated with dotted lines in Figure 1c). However, this approach is applicable only if a relevant threshold of interest on the fold change is known in advance which is not always the case in microarray experiments.

In contrast, for the two-sided criterion $p(\delta_g, 0)$ (Figure 1b), the interval of observed fold changes where the posterior probability to be differentially expressed is close to zero, is very narrow, and genes with $\bar{y}_g$ around zero can have high posterior probability. Hence this posterior probability, as well as the corresponding p-value, is difficult to interpret on its own, and in practice a joint criterion with cutoffs both on the corresponding p-value and on the fold change is often applied.

On the other hand, the tail posterior probability $p(t_g, t^{(\alpha)})$ (Figure 1a) is much less peaked around zero than $p(\delta_g, 0)$, taking small values for observed differences around zero and values close to one for large absolute differences, thus specifying groups of genes with low as well as high probability of differential expression, without relying on arbitrary threshold of interest. In addition, it varies less steeply as a function of $\bar{y}_g$ between low and high probabilities compared to $p(\delta_g, 0)$, with a large proportion of genes in between, thus allowing to choose genes with the desired level of uncertainty about their differential expression (for a theoretical comparison of gradients see Web Appendix E).

Another interesting property is the right hand tail of size α for values of the posterior probability above 0.5 under the null hypothesis (9) that helps to visualise the distribution of the tail posterior probability, in contrast with the uniform behaviour of p-values (see, e.g., Figure 3c). These two properties make the posterior probability $p(t_g, t^{(\alpha)})$ advantageous to use as a selection rule.

4. False discovery rate for tail posterior probability

In the considered objective setup, we cannot use Bayesian estimates of the false discovery rate (FDR), as proposed by Müller et al. (2006) since we do not model the alternative. Therefore, we search for a frequentist estimator of FDR. In this section we assume that the variances are the same for both conditions $\sigma_{gs}^2 = \sigma_g^2$ and have inverse gamma prior distribution $\sigma_g^{-2} \sim \Gamma(a, b)$, the same assumptions that we apply to the microarray data in Section 5. Our approach to estimate the FDR can be extended to a more general case, by approximating the distribution of w_g^2 by an inverse gamma distribution.

4.1 *Distribution of the tail posterior probability under H_0.*

In this section we treat the tail posterior probability $p(t_g, t^{(\alpha)})$ as a test statistic for testing hypothesis (2), i.e. as a function of $\bar{y}_g$.

In Section 3.1, we approximated the posterior distribution of the parameter t_g under H_0 by taking the sample difference to be zero, *i.e.* $\bar{y}_g = 0$. In this section, the marginal distribution of $\bar{y}_g$ is used to study the exact cumulative distribution function $F_0(x)$ of the tail posterior probability $p(t_g, t^{(\alpha)})$ under the null hypothesis (2) as a function of $\bar{y}_g$. We can show that the distribution function $F_0(x)$ is gene-independent and can be evaluated in practice as the empirical distribution function of expression

$$E_v[\Phi(-t^{(\alpha)} + \xi\sqrt{v/A}) + \Phi(-t^{(\alpha)} - \xi\sqrt{v/A})],$$

where the expectation is taken with respect to $v \sim \Gamma(A, 1)$, $\xi \sim t_{2A}$ and, for simplicity, with plug-in value of A, $\hat{A} = E(a \mid y_{gsr}) + \frac{1}{2}(m_1 + m_2) - 1$, where $E(a \mid y_{gsr})$ is the posterior expected value of hyperparameter a (see Web Appendix D for details).

4.2 *False discovery rate*

To estimate the frequentist expected false discovery rate (FDR) corresponding to the selection rule $p(t_g, t^{(\alpha)}) > p_{cut}$, optimal under the loss function defined in Section 3.2, we apply the approach proposed by Storey (2002):

$$FDR(p_{cut}) = \frac{\pi_0 P\{p(t_g, t^{(\alpha)}) > p_{cut} | H_0\}}{P\{p(t_g, t^{(\alpha)}) > p_{cut}\}}, \tag{10}$$

where π_0 is the probability that a gene is not differentially expressed. As shown in the previous section, the value of $P\{p(t_g, t^{(\alpha)}) > p_{cut} | H_0\} = 1 - F_0(p_{cut})$ in the numerator is gene-independent under the null hypothesis given the hyperparameters a and b and can be evaluated by numerical integration. A natural way to estimate the probability $P\{p(t_g, t^{(\alpha)}) > p_{cut}\}$ in the denominator is to take the proportion of genes whose posterior probability is above p_{cut}. To estimate π_0, we adapt the approach of Storey (2002) using the following estimate:

$$\hat{\pi}_0(\lambda) = \frac{Card\{g : g \in \{1, \ldots, n\} \,\&\, p(t_g, t^{(\alpha)}) < \lambda\}}{nF_0(\lambda)} \tag{11}$$

with small values of λ (e.g. in $[0.1, 0.2]$), where $Card\{S\}$ is the number of elements in set S. Then, we use the median of the computed $\hat{\pi}_0(\lambda)$ over a chosen range of λ as the estimate of π_0. For the simulated data (described in Section 6.2), the estimates of π_0 are shown in Table 1. There is close agreement between estimated and true values. The plots of the estimated FDR for the simulated data are given in Figure 2.

We can see that the false discovery rate is well estimated for three pairwise comparisons (details in Section 5). R code to compute the FDR estimate is available from http://www.bgx.org.uk.

[Table 1 about here.]

[Figure 2 about here.]

5. Application of tail posterior probability to microarray data

In this section we apply the tail posterior probability approach to microarray data y_{gtcr} which is described in Web Appendix A. The experimental setup has 2 factors: time points t (0, 2, and 12 hours) and conditions c (control, insulin treatment and metformin treatment) where at 0 hours only control treatment was measured, and at the two remaining time points all three conditions were measured. Thus, the data has 7 conditions, with 3 replicates r in each condition, and 22690 genes g. The question was to discover genes reacting to insulin or metformin treatment at each time point.

The model for two groups described in Section 2.1 is extended to the multiclass data:

$$y_{gtcr} \sim N(\alpha_g + \gamma_{gt} + \delta_{gtc}, \sigma_g^2), \tag{12}$$

setting $\gamma_{g0} = \delta_{gt0} = \delta_{g0c} = 0$ for model identifiability. The prior distributions are $\alpha_g, \gamma_{gt}, \delta_{gtc} \sim 1$, and following the extensive microarray literature (Gottardo et al., 2006, Smyth, 2004), we take $\sigma_g^{-2} \sim \Gamma(a, b)$. Predictive checks by Lewin et al. (2006) showed that it is an acceptable choice for modelling biological variability in microarray experiments. Care needs to be excercised when specifying the hyperprior distributions for a and b. Whilst it can be shown that the posterior distribution of a under the improper prior is proper, so that we can choose a conventional hyperprior settings $a \sim \Gamma(0.01, 0.01)$ that approximates non-informative distribution with density $f(x) = 1/x$, $x > 0$, this is not a valid choice for b since as this improper prior for b implies an improper posterior. In this case, we adapt the recommendations of Gelman (2006) to use a uniform distribution on the standard deviation: $1/\sqrt{b} \sim U[0, u]$. In the results presented, we choose the value of u to be $[\min_g(s_g^2)]^{-1/2}$ and we checked that for the considered data the results were not sensitive to this choice (for more details and justification of the prior and hyperprior choice see Web Appendix B). This model was fitted to the microarray data using WinBUGS (Spiegelhalter, Thomas and Best, 1999) (see Web Appendix F for the WinBUGS code).

The pairwise comparisons of treated to non-treated cells at each time point are considered to have relevant biological interpretation which corresponds to testing hypothesis (2) on parameters δ_{gct} in model (12), $c = 1, 2$ for insulin and for metformin treatment respectively, and $t = 1, 2$ for 2 and 12 hours. For each comparison, we calculate the tail posterior probability defined by (8) with $\alpha = 0.05$.

[Table 2 about here.]

[Figure 3 about here.]

Histograms of the posterior probability and the estimates of FDR defined in Section 4.2 (smoothed using splines) for each comparison are given in Figure 3. The first conclusion is that there is a large number of differentially expressed genes between treated and non-treated cells at 2 hours, whereas very few genes change between treated and non-treated cells at 12 hours. On the FDR plots, we can see that for the same range of values on the y-axis, the corresponding number of differentially expressed genes on the x-axes differs widely for each comparison. The number of differentially expressed genes corresponding to FDR value of 1.2% as well as the estimates of π_0 defined by (11) are given in Table 2.

For some studies, including the one considered here, we need to choose the same posterior probability cutoff for all comparisons to help biological interpretation of the study. In this case, based on both selection criteria and the lack of considerable change in gene expression in 12 hour comparisons, we choose the posterior probability cutoff to be 0.92. This corresponds to 1372 genes and $FDR = 1\%$ in comparison of insulin to control at 2 hours.

6. Extension of tail posterior probability to the analysis of multiclass data

6.1 *Joint tail posterior probability*

Now we consider the problem of testing a hypothesis involving several differential parameters of interest, which we call a *compound null hypothesis*, a situation that often arises in multiclass microarray experiments. For instance, in the considered microarray data we discovered two comparisons where there is a noticeable change in gene expression: for each type of treated cells to non-treated cells at 2 hours (Section 5), and the next question is to find genes differentially expressed in both comparisons. Hence, we need to test the null hypothesis for two (or more comparisons) for each gene against the specific alternative (we omit index $t = 1$ for simplicity):

$$H_0^{(g)}: \quad \delta_{g1} = 0 \;\&\; \delta_{g2} = 0 \quad \text{versus} \quad H_1^{(g)}: \quad \delta_{g1} \neq 0 \;\&\; \delta_{g2} \neq 0. \tag{13}$$

Simply rejecting the null hypothesis in ANOVA settings corresponds to a different alternative from the one required. To rank genes with respect to the degree of their differential expression in both comparisons, we use an extension of the tail posterior probability, the *joint tail posterior probability*, which is motivated by the interpretation of the tail posterior probability as a posterior expected discrepancy between the hypotheses:

$$p_g^J \;=\; E(I\{|t_{g1}| > t^{(\alpha)}\}I\{|t_{g2}| > t^{(\alpha)}\} \mid \mathbf{y}) = P\{|t_{g1}| > t^{(\alpha)} \;\&\; |t_{g2}| > t^{(\alpha)} \mid \mathbf{y}\},$$

where $t_{gc} = \delta_{gc}/w_{gc}$ is the standardised difference defined in Section 3.1. The joint posterior probability is appropriate in the multiclass settings where the standardised differences t_{g1} and t_{g2} are dependent *a posteriori*, leading to a fewer number of false positives. In the considered microarray example, the standardised differences are dependent due to sharing the same variance parameter $w_{gc}^2 = 2\sigma_g^2/3$ (see Web Figure 5 for histograms of posterior correlation between the differences δ_{g11} and δ_{g12} and between the standardised differences t_{g11} and t_{g12} for the H2Kb data).

We compare this method to a common solution to this problem used in practice which is to select a gene list for each comparison separately using the same cutoff, and find genes common to the lists. It can be formalised by

$$p_g^{min} = \min\left(P\{|t_{g1}| > t^{(\alpha)} \mid \mathbf{y}\}, P\{|t_{g2}| > t^{(\alpha)} \mid \mathbf{y}\}\right). \tag{14}$$

6.2 *Simulated data*

To illustrate our methodology, we use a data set simulated from the model (12), with $a = 2$, $b = 0.05$, $n = 2000$ and 3 replicates. Non-zero values of the linear parameters are simulated from $(-1)^{Bern(0.5)}N(0.9, 0.5^2)$, with no non-zero values for γ_{g1}, 20 non-zero values for γ_{g2}, 100 non-zero values for δ_{g11}, 200 non-zero values for δ_{g12}, 600 non-zero values for δ_{g21} and 2 non-zero values for δ_{g22}. Out of 200 non-zero values of δ_{g12}, 36 are chosen to be the same of those of δ_{g11}. The purpose is to create some commonality of effects between the two conditions observed in our case study.

6.3 *Performance of joint tail posterior probability*

To compare the performance of the two rules, we can study the number of false negatives and false positives for these two rules on the simulated data (Web Figure 6). It shows that the lists of differentially expressed genes identified by joint tail posterior probability have fewer number of false positives, and about the same number of false negatives, thus leading to a rule with consistently higher power.

Applying the joint tail posterior probability to the microarray data with the cutoff 0.92, 280 genes are declared to be differentially expressed in both comparisons, and combining the individual gene lists using criterion (14) adds extra 47 genes to the list which is explained by the relation $p_g^{min} \geq p_g^J$.

7. Discussion

In this paper, we introduced a new measure of differential expression based on the posterior distribution of standardised difference, the tail posterior probability, which is an optimal decision rule with respect to a specific loss function and which can also be interpreted as a discrepancy measure for model rejection considered in Bernardo & Smith (2004) in the objective settings. We compared the tail posterior probability with other posterior measures of differential expression and found that it possesses good properties, such as smooth variation as a function of $\bar{y}_g$ and that it avoids specifying *a priori* a threshold on the fold change. Moreover, using the interpretation of the tail posterior probability as a test statistic (in the frequentist sense), we derived its distribution under the null hypothesis which enables us to propose a frequentist estimate of the false discovery rate associated with this measure. Even though our results were discussed in the context of gene expression experiments, the hierarchical model formulated in (1) and (3) is generic to a large number of situations, e.g. toxicology experiments, drug testing and meta-analysis (see discussion and references in Breslow, 1990). We are not aware that tail posterior probabilities have been used formally in such applications, but these setups would certainly allow their use. In general, the tail posterior probability approach could be useful in medical applications where a point null hypothesis is inappropriate and thus an interval null hypothesis must be considered; however, the length of the null interval is usually unknown and needs to be identified, for example, in studying "bioequivalence" of two drugs (Breslow, 1990). Besides medical applications, rules based on posterior probabilities have been used in epidemiology in the disease mapping context. In particular, maps of the posterior probability that the relative risk is greater than one have been commonly proposed as a tool to highlight areas of greater risk. Here the rule is akin to the $p(d_g, 0)$ rule since 1 is the value of the relative risk under the null hypothesis, and a study of the performance of such rules has recently been performed (Richardson et al., 2004). The extension of tail posterior probabilities to this setup is an interesting avenue for further research.

Our motivation for studying tail posterior probabilities was that besides their obvious use in simple differential expression setups, such flexible rules are particularly adapted to testing compound hypotheses that arise in generalised ANOVA settings. We showed that tail posterior probability selection rule can be naturally extended to multiclass settings using its interpretation as a discrepancy measure and that it yields a lower number of false positives compared to the traditional approach of combining individual gene lists wherever there is dependence between the different comparisons of interest.

In general, the rich output produced by a Bayesian analysis of multiclass experiments allows a number of decision rules to be investigated. As a first step, we showed equivalence between frequentist testing based on the marginal distribution of the sample difference and selecting gene lists using posterior probabilities estimated in the corresponding Bayesian model. Characterising general properties and comparing decision rules based on tail posterior probabilities in more general setups than the hierarchical model considered in this paper, are interesting research topics for future work.

8. Supplementary materials

Web Appendices and Figures referenced in Sections 3-6 are available under the Paper Information link at the Biometrics website www.tibs.org/biometrics.

Acknowledgements

Natalia Bochkina's work was funded by a Wellcome Trust Cardio - Vascular grant 066780/Z/01/Z. The authors would like to thank Ulrika Andersson and

David Carling who generated the data, their colleagues Anne-Mette Hein, Alex Lewin and Maria de Iorio for stimulating discussion on the statistical aspects, Susie Bayarri and Judith Rousseau for helpful suggestions and the BAIR group (BAIR project, www.bair.org.uk) for sharing their biological insights. We are also grateful to the referee and the associate editor for their useful comments that helped us to improve the manuscript.

References

Baldi, P. and Long, A. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**(6), 509–519.

Bernardo, J.M. and Smith, A.F.M. (2004) Bayesian Theory. Wiley, NY.

Breslow, N. (1990) Biostatistics and Bayes. *Statistical Science* **5**, 269–198.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin D.B. (2004) Bayesian Data Analysis (2nd edition). Series *Textbooks in Statistical Science*, Chapman and Hall/CRC.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**(3), 515–533.

Gottardo, R., Raftery, A.E., Yee Yeung, K., and Bumgarner, R.E. (2006) Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples. *Biometrics* **62**(1), 1–9.

Jeffreys, H. (1961) Theory of Probability. Oxford University Press, London.

Lewin, A., Richardson, S., Marshall, C., Glazier, A., and Aitman, T. (2006) Bayesian Modeling of Differential Gene Expression. *Biometrics* **62**(1), 10–18.

McLachlan, G.J., Do, K.-A., and Ambroise, C. (2004) Analyzing Microarray Gene Expression Data. Wiley, New York.

Müller, P., Parmigiani, G., and Rice, K. (2006) FDR and Bayesian decision rules, to appear in *Bayesian statistics* **8**, Oxford University press.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays. *Journal of the American Statistical Association* **99**, 990–1001.

Richardson, S., Thomson, A., Best, N., and Elliott, P. (2004) Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives* **112**, 1016–1025.

Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**: No. 1, Article 3.

Spiegelhalter, D. J., Thomas, A., and Best, N. (1999). WinBUGSVersion 1.4 User Manual. MRC Biostatistics Unit. Software available at http://www.mrcbsu.cam.ac.uk/bugs/winbugs/contents.shtml.

Storey, J. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

Appendix A

Conditions for equivalence of Bayesian and frequentist inference

In this section we state and prove the theorem generalising Theorem 1. To simplify the notation, we omit the gene-related index g. Theorem 1 can be obtained from Theorem 2 by taking $G(x-\delta \mid \theta) = \Phi((x-\delta)/\sigma)$ with $\theta = \sigma^2$, and $y_a = s^2$ is the sample variance.

Theorem 2. *Assume that $\bar{y}$ is a sufficient statistic for δ, and $\bar{y} \mid \delta, \theta \sim G(x - \delta \mid \theta)$, where $G(x \mid \theta)$ is a distribution function with a vector of parameters θ, and we want to test whether δ is zero. Denote the vector of auxiliary statistics with respect to statistic $\bar{y}$ by y_a. We assume Bayesian settings with the (improper) uniform prior distribution for δ, and that the prior distribution of θ does not depend on δ.*

Then, $P\{\delta > 0 \mid y_{sr}\} = 1 - P\{\eta > \bar{y}\}$ where the random variable η has the same distribution as $\bar{y}$ given $\delta = 0$ and y_a, and thus $P\{\eta > \bar{y}\}$ with the observed value of $\bar{y}$ is a p-value. Also, distribution of $P\{\delta > 0 \mid y_{sr}\}$ is uniform under the null hypothesis.

Proof. (Theorem 2) . By the definition of the sufficient statistic, the likelihood $f(y_{sr} \mid \delta, \theta)$ can be represented as a product of two density functions: $f(\bar{y} \mid \delta)$ and $f(y_a \mid \theta)$, where $y_a = S(y_{sr})$, $S : \mathbb{R}^{m_1+m_2} \to \mathbb{S} \subset \mathbb{R}^{m_1+m_2-1}$, is an auxiliary statistic with respect to $\bar{y}$. Therefore, the marginal distribution of y_{sr} and the joint posterior distribution of θ and δ also factorise:

$$\begin{aligned}
f(y_{sr}) &= \int_{x,z} f(y_{sr} \mid \theta = z, \delta = x) f_\theta(z) f_\delta(x) dx dz \\
&= \int_z f(y_a \mid \theta = z) \int_x f(\bar{y} \mid \theta = z, \delta = x) f_\theta(z) f_\delta(x) dx dz \\
&= \int_z f(y_a \mid \theta = z) f_\theta(z) dz \int_x g(\bar{y} - x \mid \theta = z) dx = f(y_a) 1(\bar{y}) \\
f(\delta, \theta \mid y_{sr}) &= \frac{f(y_{sr} \mid \delta, \theta) f(\delta) f(\theta)}{f(y_{sr})} = \frac{f(\bar{y} \mid \delta, \theta) f(\delta)}{f(\bar{y})} \frac{f(y_a \mid \theta) f(\theta)}{f(y_a)} \\
&= g(\bar{y} - \delta \mid \theta) f(\theta \mid y_a),
\end{aligned}$$

where $g(x \mid \theta)$ is the density function of distribution $G(x \mid \theta)$. This implies that the posterior distributions of parameters δ and θ are:

$$\begin{aligned}
F(\delta \mid y_{sr}) &= \int_\theta G(\delta - \bar{y} \mid \theta) f(\theta \mid y_a) d\theta, \\
f(\theta \mid y_{sr}) &= f(\theta \mid y_a).
\end{aligned}$$

Note that the assumptions of the uniform prior for δ and independence of the prior for θ of δ are crucial here since they allow factorisation of marginal and posterior distributions.

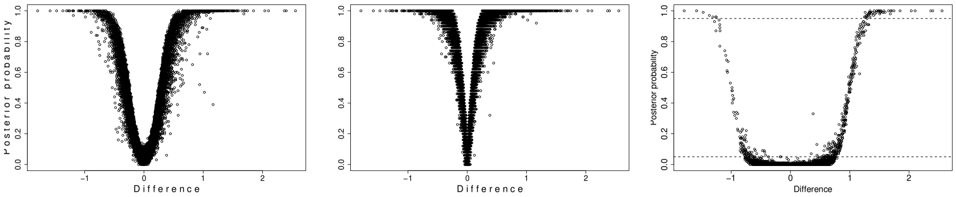
Thus we see that the posterior distribution of θ does not depend on $\bar{y}$, and the distribution function of η is $F_\eta(x \mid y_a) = P\{\bar{y} < x \mid \delta = 0, y_a\} = \int G(x - d \mid \theta) f(\theta \mid y_a) d\theta$.

Therefore, the posterior probability of interest can be written as follows:

$$P\{\delta > 0 \mid y_{sr}\} = \int G(-\bar{y} \mid \theta) f(\theta \mid y_a) d\theta = 1 - F_\eta(\bar{y}).$$

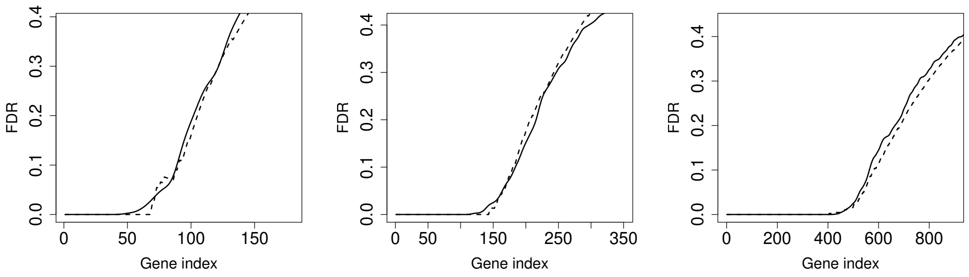
Also, distribution of $P\{\delta > 0 \mid y_{sr}\}$ is uniform under the null hypothesis since $\bar{y}$ and η have the same distributions. Thus, Theorem 2 is proved.

List of Figures



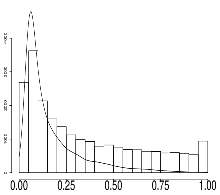
(a) $p(t_g, t^{(\alpha)})$ (b) $p(\delta_g, 0)$ (c) $p(\delta_g, \log_2 2)$

Figure 1.

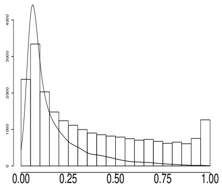


(a) Comparison 1 (b) Comparison 2 (c) Comparison 3

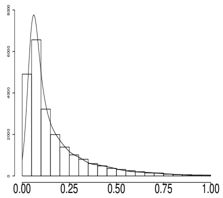
Figure 2.



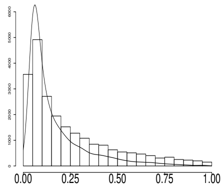
(a) Insulin vs Control 2hrs



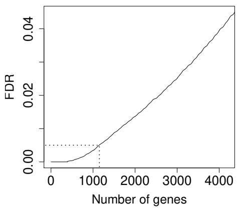
(b) Metformin vs Control 2hrs



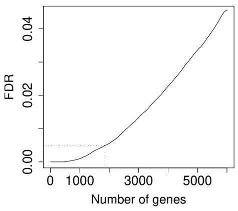
(c) Insulin vs Control 12hrs



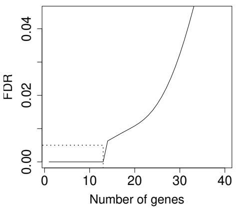
(d) Metformin vs Control 12hrs



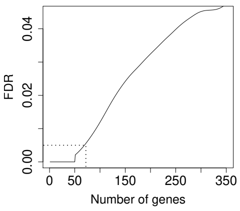
(e) Insulin vs Control 2hrs



(f) Metformin vs Control 2hrs



(g) Insulin vs Control 12hrs



(h) Metformin vs Control 12hrs

Figure 3.

Table 1

Estimate of π_0 for simulated data in comparison 1 (parameter δ_{g11}), comparison 2 (δ_{g12}), comparison 3 (δ_{g21}) and comparison 4 (δ_{g22}).

Comparison	1	2	3	4
Estimated π_0	0.953	0.907	0.718	1.000
True π_0	0.950	0.900	0.700	0.999

Table 2

Number of differentially expressed genes and the posterior probability cutoff corresponding to a gene list with estimated $FDR = 1.2\%$ in H2Kb data.

Comparison	Insulin vs Control 2hrs	Insulin vs Control 12hrs	Metformin vs Control 2hrs	Metformin vs Control 12hrs
Estimated π_0	0.61	0.99	0.56	0.80
Number of genes	1475	13	1854	72
Posterior probability cutoff p_{cut}	0.91	1.00	0.92	0.99