Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 5, *pp.* 905–925

A Bayesian discovery procedure

Michele Guindani,

University of New Mexico, Albuquerque, USA

Peter Müller

University of Texas M. D. Anderson Cancer Center, Houston, USA

and Song Zhang

University of Texas Southwestern Medical Center, Dallas, USA

[Received March 2008. Revised February 2009]

Summary. We discuss a Bayesian discovery procedure for multiple-comparison problems. We show that, under a coherent decision theoretic framework, a loss function combining true positive and false positive counts leads to a decision rule that is based on a threshold of the posterior probability of the alternative. Under a semiparametric model for the data, we show that the Bayes rule can be approximated by the optimal discovery procedure, which was recently introduced by Storey. Improving the approximation leads us to a Bayesian discovery procedure, which exploits the multiple shrinkage in clusters that are implied by the assumed non-parametric model. We compare the Bayesian discovery procedure and the optimal discovery procedure estimates in a simple simulation study and in an assessment of differential gene expression based on microarray data from tumour samples. We extend the setting of the optimal discovery procedure by discussing modifications of the loss function that lead to different single-thresholding statistics. Finally, we provide an application of the previous arguments to dependent (spatial) data.

Keywords: Bayes optimal rule; False discovery rate; Loss function; Multiple comparison

1. Introduction

Various approaches have been introduced in the recent literature to address the multiple-comparison problem. Most focus on controlling some error rate. For example, the control of the familywise error rate guarantees a bound on the probability of a false rejection among all tests. Benjamini and Hochberg (1995) developed a simple procedure, based on the ordered p-values, that controls the false discovery rate (FDR), which is defined as the expected proportion of rejected null hypotheses which are erroneously rejected. A decision theoretic approach to the multiple-comparison problem requires the explicit statement of a loss function, which weights the relative importance of the various outcomes according to the preferences and inferential focus of the investigators. Cohen and Sackrowitz (2007) proved the inadmissibility of the Benjamini and Hochberg procedure under any loss that is a linear combination of false discoveries and false acceptances and under several sampling models, including the general one-parameter exponential family. Müller *et al.* (2004, 2007) undertook a decision theoretic approach to multiple testing and discussed several loss functions that lead to the use of FDR-based rules.

Address for correspondence: Michele Guindani, Department of Mathematics and Statistics, MSC03 2150, University of New Mexico, Albuquerque, NM 87131-0001, USA.
E-mail: michele@stat.unm.edu

 1369–7412/09/71905

More recently, Bogdan *et al.* (2008) compared the Benjamini–Hochberg procedure with several Bayesian rules for multiple testing. They showed that, whenever the proportion of true null hypotheses is small, the misclassification error of the Benjamini–Hochberg procedure is close to optimal, in the sense of matching a Bayesian oracle. This property is shown to be shared by some of the Bayesian procedures that they considered. In addition, through simulations, they showed that Bayes rules generally perform better for large or moderate proportions.

The general multiple-comparison problem is stated as follows. Assume that we observe data sets $\mathbf{x}_1, \ldots, \mathbf{x}_m$, where $\mathbf{x}_i = \{x_{1i}, \ldots, x_{n_i i}\}$, and for each $\mathbf{x}_i$ we consider a test of a null hypothesis H_{0i}. Often the data are reduced to a statistic z_i with $z_i \sim f(z_i; \mu_i)$, for some distribution f, indexed by an unknown parameter μ_i, $i = 1, \ldots, m$. Assume that we wish to test $H_{0i} : \mu_i \in A$ *versus* $H_{1i} : \mu_i \notin A$ for $i = 1, \ldots, m$. We discuss the multiple-comparison problem in a Bayesian decision theoretic framework. A Bayesian decision problem is characterized by a sampling model, $f(z_i; \mu_i)$ in this case, a prior and a loss function which represents the utility preferences that are associated with possible outcomes.

Specifically, we use a prior model that includes a random-effects distribution G for the μ_i. Instead of a parametric model for G, we consider G as an unknown random probability measure. A prior probability model for a random probability measure is known as a non-parametric (NP) Bayesian model. We assume a Dirichlet process (DP) prior, which is one of the most popular NP Bayesian models.

For sufficiently large m and a small total mass parameter of the DP prior, the posterior random probability measure can be approximated by the empirical distribution of the maximum likelihood estimates $\hat{\mu}_i$. The result is an approximate Bayes rule that is closely related to Storey's optimal discovery procedure (ODP) (Storey, 2007).

The ODP is based on a thresholding statistic,

$$S_{\mathrm{ODP}}(z_i) = \sum_{\mu_j \notin A} f(z_i; \mu_j) \Big/ \sum_{\mu_j \in A} f(z_i; \mu_j). \tag{1}$$

The null hypothesis H_{0i} is rejected if $S_{\mathrm{ODP}}(z_i) \geqslant \lambda$, for some $0 \leqslant \lambda < \infty$. Let $d_i^{\mathrm{ODP}} = I\{S_{\mathrm{ODP}}(z_i) \geqslant \lambda\}$. Storey proved that d^{ODP} maximizes the expected number of true positive results among all procedures with equal or smaller expected number of false positive results. For a point null hypothesis $A = \{0\}$, the test reduces to thresholding

$$S_{\mathrm{ODP}}(z_i) = \sum_{\mu_j \notin A} f(z_i; \mu_j) \Big/ f(z_i; 0).$$

As stated, the threshold function S_{ODP} involves the unknown parameters μ_j, $j = 1, \ldots, m$. In practice, $S_{\mathrm{ODP}}(\cdot)$ must be estimated. For a point null hypothesis, i.e. $A = \{0\}$, the ODP is evaluated as

$$\hat{S}_{\mathrm{ODP}}(z) = \sum_{j=1}^{m} f(z; \hat{\mu}_j) \Big/ f(z; 0), \tag{2}$$

where $\hat{\mu}_j$ is a point estimate for μ_j (e.g. the maximum likelihood estimate (MLE)). It is shown that the performance of $\hat{S}_{\mathrm{ODP}}$ is comparable with the theoretically ODP based on S_{ODP}.

For general A the ODP proceeds with the thresholding function

$$\hat{S}_{\mathrm{ODP}}(z) = \sum_{j=1}^{m} f(z; \hat{\mu}_j) \Big/ \sum_{j=1}^{m} w_j f(z; \hat{\mu}_j), \tag{3}$$

where w_j are suitable weights, chosen to estimate the true status of each hypothesis. For example, $w_j = 1$ for all comparisons that are estimated (by some preliminary test) to be null,

and $w_j = 0$ otherwise. Storey *et al.* (2007) showed that function (3) outperforms many procedures that are commonly used in testing a large number of genes for differential expression.

We show that the ODP can be recognized as an approximate Bayes rule under the proposed NP prior model for μ_i. This result is in accordance with Ferguson's (1973) observation that NP Bayesian inference often yields results that are comparable with corresponding classical inference. The expectation in Storey's optimality statement for d^{ODP} is under the (frequentist) repeated sampling model. The expectation in the Bayes rule is under the posterior distribution for a given outcome z. Maximization of the conditional expectation for each z implies maximization of the marginal expectation across repeated sampling. A similar argument can be made about the constraint on the expected number of false positive results. A bound on the conditional expectation, conditional on each outcome z, implies the same bound on the marginal expectation. We conclude that the Bayes rule under the NP prior approximately satisfies the optimality property of the ODP procedure. By the same arguments we show that thresholding the marginal posterior probability amounts to controlling the positive FDR (Storey, 2002; Storey and Tibshirani, 2003).

Once we have established the utility function and the probability model leading to the ODP we can proceed with generalizations in two directions. First, we shall consider variations of the ODP to improve the approximation. We show that the resulting rules lead to small improvement in inference. More importantly, once we recognize the ODP as a Bayes rule we can modify the procedure to adapt to variations in the loss function. We provide specific examples, including a study of exceedance regions of a spatial process and the detection of neurodegenerative patterns in magnetic resonance imaging (MRI) scans.

DP priors in the context of multiple-hypotheses testing have been considered before by Gopalan and Berry (1993). More recently, Dahl and Newton (2007) proposed a DP mixture model (called BEMMA) for testing correlated hypotheses and showed that the induced clustering information leads to an improved testing power. Unrelated to a discussion of Storey's ODP, Bogdan *et al.* (2008) proposed the same semiparametric model as the model that we introduce in Section 3 and used it for the comparison of misclassification rates among several competing models. The distinction between the previous approaches and ours is that here the DP prior is part of a decision theoretic set-up. Besides, both Gopalan and Berry (1993) and Dahl and Newton (2007) restricted inference to hypotheses concerning the configuration of ties between the parameters of interest. A Bayesian implementation of the ODP procedure has recently also been considered by Cao *et al.* (2009). They developed a parametric Bayes hierarchical mixture model that achieves shrinkage of the gene-specific sampling variances. The resulting posterior means are then plugged in the ODP statistic and this is shown to outperform the original ODP and many widely used competing procedures. The discussion in Cao *et al.* (2009) is purely empirical, without reference to a decision theoretic set-up.

The format of the paper is as follows. In Section 2, we introduce the decision problem and the resulting Bayes rule. In Section 3 we introduce the NP probability model and we detail the algorithm for posterior inference. In Section 4 we finally discuss the interpretation of the ODP as an approximate Bayes rule and introduce the corresponding Bayesian discovery procedure (BDP) statistics. In Section 5 we compare the behaviour of the ODP and the BDP with simulated data and a microarray data set. We show that the BDP provides at least some improvement over the frequentist optimal procedure. In Section 6 we discuss some extensions of the previous settings for multigroup experiments, different loss functions and different inferential purposes. In particular, we consider a spatially dependent NP model and apply the Bayesian multicomparison decision rule to a simplified MRI data set.

2. The decision problem

For a formal definition of the multiple-comparison problem we need some notation and minimal assumptions on the sampling model. Assume that the data are $z_i|\mu_i \sim f(z_i;\mu_i)$, independently across i, $i=1,\ldots,m$. The competing hypotheses are formalized as

$$H_{0i}:\mu_i \in A \quad \textit{versus} \quad H_{1i}:\mu_i \notin A,$$

using, for example, $A=(-\varepsilon,\varepsilon)$ or $A=\{0\}$. Let G denote the distribution of μ_i, which is obtained by marginalizing over the two competing hypotheses, and let π_0 denote the prior probability of the null hypothesis, i.e. $\pi_0 \equiv G(A)=p(H_{0i})$. Let $G(\mu|A) \propto G(\mu)\, I(\mu \in A)$ denote G conditional on A. The model can be written as a mixture prior,

$$p(\mu_i|G)=\pi_0\, G(\mu_i|A)+(1-\pi_0)\, G(\mu_i|A^{\mathrm{c}}),$$

where A^{c} denote the complement set of A. Alternatively, the model can be defined as a hierarchical model by means of a latent indicator parameter $r_i \in \{0,1\}$, which is interpreted as the (unknown) truth of the ith comparison.

$$p(\mu_i|r_i)=\begin{cases} G(\mu_i|A) & \text{if } r_i=0, \\ G(\mu_i|A^{\mathrm{c}}) & \text{if } r_i=1, \end{cases} \text{ and } \Pr(r_i=0)=\pi_0. \tag{4}$$

We shall use $z=(z_1,\ldots,z_m)$ and $\theta=(G,r_i,\mu_i,i=1,\ldots,m)$ to refer generically to the data and parameters in model (4).

In addition to a probability model, a Bayesian decision theoretic approach is characterized by a set of actions (decisions) and a loss function corresponding to all possible outcomes of the experiment. Let $d_i \in \{0,1\}$ denote the decision for the ith hypothesis, with $d_i=1$ indicating a decision against H_{0i}, and let $d=(d_1,\ldots,d_m)$. To define an optimal rule for d_i we introduce a loss function $L(d,\theta,z)$. The optimal rule $d_i^*(z)$ is defined by minimizing L in expectation with respect to the posterior model $p(\theta|z)$. Formally,

$$d^*=\arg\min_d \left\{ \int L(d,\theta,z)\, p(\theta|z)\, \mathrm{d}\theta \right\}.$$

We use a loss function that combines the true positive count $\mathrm{TP}=\Sigma\, d_i r_i$ and false positive count $\mathrm{FP}=\Sigma\, d_i(1-r_i)$,

$$L(d,\theta,z)=-\sum d_i r_i+\lambda \sum d_i(1-r_i)=-\mathrm{TP}+\lambda \mathrm{FP}. \tag{5}$$

The loss (5) can be interpreted as the Lagrangian for maximizing TP subject to a given bound on FP. It is a multiple-comparison equivalent of the loss function underlying the Neyman–Pearson paradigm for a simple test.

Let $v_i=E(r_i|z)$ denote the marginal posterior probability for the ith alternative hypothesis. It is straightforward to show that the optimal rule under loss (5) is a threshold on v_i,

$$d_i^*=I(v_i>t)=I\left(\frac{v_i}{1-v_i}>t_2\right). \tag{6}$$

Alternatively, the threshold on v_i can be written as a threshold on the posterior odds $v_i/(1-v_i)$. The statement is true for any probability model (subject only to the stated quantities having meaningful interpretations). Moreover rules based on thresholding the marginal posterior probability imply control of the frequentist positive FDR (Storey, 2002; Storey and Tibshirani, 2003).

Proposition 1. Consider m hypothesis tests $H_{0i}:\mu_i \in A$ *versus* $H_{1i}:\mu_i \in A^{\mathrm{c}}$, data $z_1,\ldots,z_m$, where $z_i|\mu_i \sim f(z_i;\mu_i)$, independently across i, $i=1,\ldots,m$, and a prior probability model $p(\mu_i|G)=$

$\pi_0 G(\mu_i|A) + (1-\pi_0) G(\mu_i|A^c)$ for some distribution G and probability $\pi_0 = p(H_{0i}) = p(r_i = 0)$. Let the rejection region be determined by equation (6), i.e. $d_i^* = I(v_i > t)$. Let $D = \Sigma_{i=1}^m d_i$. Then,

$$\text{pFDR} = E\left(\frac{\text{FP}}{D} | D > 0\right) < 1 - t.$$

For a proof see Appendix A.

The only substantial assumption in proposition 1 is that d_i^* is a threshold on the gene-specific posterior probabilities of differential expression. Bogdan *et al.* (2008) proved a similar result for loss functions that are a linear combination of false negative and false positive counts.

Rule (6) is different from rules that are based on the local FDR. The local FDR is defined as the posterior probability of the null hypothesis given that we have observed a certain value z_i, and given assumed known sampling models under the null and alternative hypotheses (Efron *et al.*, 2001). In contrast, v_i is defined conditionally on the observed values of z across all tests. Hence, the local FDR provides a measure of significance that is local to z_i, whereas v_i is a global measure of significance.

3. A non-parametric Bayes decision rule

3.1. A semiparametric Bayesian model

We complete the sampling model (4) with a prior model for G. Prior probability models for unknown distributions, G in this case, are traditionally known as NP Bayesian models. One of the most commonly used NP Bayesian priors is the DP model. We write $G \sim \text{DP}(G^*, \alpha)$ to indicate a DP for a random probability measure G. See Ferguson (1973, 1974) for a definition and important properties of the DP model. The model requires the specification of two parameters: the base measure G^* and the total mass parameter α. The base measure G^* is the prior mean, $E(G) = G^*$. The total mass parameter determines, among other important properties, the variation of the random measure around the prior mean. Small values of α imply high uncertainty. In the following discussion we exploit two key properties of the DP. A random measure G with DP prior is almost surely discrete. This allows us to write G as a mixture of point masses, $G = \Sigma w_h \delta_{m_h}$. Another important property is the conjugate nature of the DP prior under random sampling. Assume that $\mu_i \sim G$, $i = 1, \ldots, m$, are an independent and identically distributed sample from a random measure G with DP prior, $p(G) = \text{DP}(G^*, \alpha)$. Then, the posterior probability model is $p(G|\mu) = \text{DP}(G_1^*, \alpha + m)$, for $G_1^* \propto \alpha G^* + m F_m$. Here, $F_m = (1/m)\Sigma \delta_{\mu_i}(\cdot)$ is the empirical distribution of the realized μ_is.

We use a DP prior on G to complete model (4)

$$\mu_i | G \sim G \qquad\qquad G \sim \text{DP}(G^*, \alpha). \tag{7}$$

Model (7) implies that the prior for the null hypothesis $p_0 = G(A)$ is beta, $p_0 \sim \text{Be}[\alpha G^*(A), \alpha\{1 - G^*(A)\}]$.

3.2. A semiparametric Bayes rule for multiple testing

Many approaches have been proposed in the literature to implement posterior Monte Carlo simulation for DP mixture models. See, for example, Neal (2000) for a review. We outline how these methods can be adapted to compute the posterior probabilities v_i.

Let z_i, $i = 1, \ldots, m$, denote the observed data (or a summary statistic) for test i. We assume that

$$z_i | \mu_i \overset{\text{ind}}{\sim} f(z_i; \mu_i), \qquad\qquad i = 1, \ldots, m. \tag{8}$$

For example, $f(z_i;\mu_i) \equiv N(z_i;\mu_i,\sigma)$, which is a normal distribution with mean μ_i and variance σ^2. We complete the model with an NP prior

$$\mu_i | G \overset{\text{IID}}{\sim} G \qquad G \sim \text{DP}(G^*, \alpha) \text{ with } G^*(\cdot) \sim \pi_0 h_A(\cdot) + (1-\pi_0) h_{A^c}(\cdot), \tag{9}$$

where $h_A(\cdot)$ and $h_{A^c}(\cdot)$ are distributions with support respectively on A and A^c. Equations (8) and (9) define a DP mixture model where G^* itself is a mixture of two terms. Unrelated to a discussion of the ODP, Bogdan *et al.* (2008) proposed the same model for multiple-comparison problems. They compared misclassification rates by using inference based on pFDR and the FDR.

Algorithms for posterior Monte Carlo simulation in DP mixture models can easily be modified to adapt to the mixture in G^*. We shall focus on the case $A = \{0\}$ and outline the necessary changes for general A. We set $h_A(\cdot) = \delta_0(\cdot)$, i.e. a point mass at 0. Also, we choose $h_{A^c}(\cdot)$ to be continuous, e.g. $N(0, \sigma^2)$, and we shall denote it simply by $h(\cdot)$. The almost sure discrete nature of a DP random probability measure implies a positive probability for ties in a sample from G. The ties naturally define a partition of observations into clusters with common values μ_j. We introduce latent cluster membership indicators s_i to describe this partition by $s_i = s_k$ if $\mu_i = \mu_k$. We reserve the label $s_i = 1$ for the null distribution, i.e. we set $s_i = 1$ if and only if $\mu_i = 0$. Let z_{-i} and s_{-i} denote the set of observations and the indicators excluding the ith. Also, let L be the number of clusters that are defined by ties (an unmatched single observation counts as a singleton cluster), m_s be the size of cluster s and $m_{-i,s}$ be the size of cluster s without observation i. Finally, let $\Gamma_s = \{i : s_i = s\}$ denote the sth cluster, and let $\mu_s^* = \mu_i$, $i \in \Gamma_s$, denote the common value of μ_i for cluster s. Then, the ith observation falls in one of the existing clusters or forms a new cluster according to the following modified Pólya urn scheme:

$$P(s_i = s | s_{-i}, z) \propto \begin{cases} \dfrac{m_{-i,1} + \pi_0 \alpha}{m - 1 + \alpha} f(z_i | \mu_{s_i}^* = 0) & \text{if } s = 1, \\ \dfrac{m_{-i,s}}{m - 1 + \alpha} \displaystyle\int f(z_i | \mu_s^*) h(\mu_s^* | z_{-i,s}) \, \mathrm{d}\mu_s^*, & \text{if } 2 \leqslant s \leqslant L, \\ \dfrac{(1-\pi_0)\alpha}{m - 1 + \alpha} \displaystyle\int f(z_i | \mu_s^*) h(\mu_s^*) \, \mathrm{d}\mu_s^*, & \text{if } s = L + 1. \end{cases}$$

Here, $h(\mu_s^* | z_{-i,s})$ denotes the posterior distribution of μ_s^* based on the prior $h(\cdot)$ and all observations z_h, $h \in \Gamma_s / \{i\}$. Note that the posterior of μ_1^* given all the observations in cluster 1 is still a point mass $\delta_0(\cdot)$. We described the algorithm for a particular choice of A and G^*. But it can be easily extended to more general A and G^*. In the general case, clusters are formed either by samples from $h_A(\cdot)$ or $h_{A^c}(\cdot)$. The algorithm is greatly simplified when A is an interval, $h_A(\mu) \propto h(\mu) I_A(\mu)$, $h_{A^c}(\mu) \propto h(\mu) I_{A^c}(\mu)$, for some continuous distribution $h(\mu)$ and $\pi_0 = G^*(A)$. Then, equations (8) and (9) describe a traditional DP mixture of normals model with Gaussian base measure G^*. The usual Pólya urn scheme for DP mixtures may be used.

Once we have a posterior Monte Carlo sample of the model parameters it is easy to evaluate the decision rule. Assume that we have B posterior Monte Carlo samples of random partitions, $s^{(b)} = (s_1^{(b)}, \ldots, s_m^{(b)})$, $b = 1, \ldots, B$. We evaluate $v_i = E(r_i \,|\, z)$ numerically as $\bar{v}_i = 1 - \Sigma_{b=1}^{B} I(s_i^{(b)} = 1)/B$. We evaluate d_i^* by substituting $\bar{v}_i$ in equation (6).

$$S_{\text{NP}} = I\left(\frac{\bar{v}_i}{1 - \bar{v}_i} > t \right). \tag{10}$$

In the next section we show that under a DP prior S_{NP} can be approximated by a single threshold statistic similar to expression (1).

4. The Bayesian discovery procedure

4.1. The optimal discovery procedure as approximate Bayes rule

We show that, under a DP prior model, $d^* \approx d^{\mathrm{ODP}}$, i.e. the ODP is an approximate Bayes rule. We start by writing the marginal posterior probability as expectation of conditional posterior probabilities,

$$v_i = E\{p(r_i = 1|G, z)|z\} = E\left\{ \frac{\int_{A^c} f(z_i;\mu)\,\mathrm{d}G(\mu)}{\int_{A\cup A^c} f(z_i;\mu)\,\mathrm{d}G(\mu)} |z \right\},$$

and proceed with an approximation of the conditional posterior distribution $p(G|z)$. The conjugate nature of the DP prior under random sampling implies that $E(G|\mu, z) = G_1^* \propto \alpha G^* + \Sigma\delta_{\mu_i}$. Recall that $F_m \propto \Sigma\delta_{\hat{\mu}_i}$ is the empirical distribution of the MLEs $\hat{\mu}_i$. For large m, small α and an informative sampling model $f(z_i;\mu)$, $E(G|z) \approx F_m$. Further, for large m, the uncertainty of the posterior DP, $p(G|\mu, z) = \mathrm{DP}(G_1^*, m+\alpha)$, is negligible, allowing us to approximate the posterior on the random probability measure G with a degenerate distribution at F_m, $p(G|z) \approx \delta_{F_m}(G)$, i.e. $G \approx (1/m)\Sigma\delta_{\hat{\mu}_i}$. Therefore,

$$v_i \approx \frac{\int_{A^c} f(z;\,\mu)\,\mathrm{d}F_m(\mu)}{\int f(z;\mu)\,\mathrm{d}F_m(\mu)} = \frac{\sum_{\hat{\mu}_j\in A^c} f(z_i;\hat{\mu}_j)}{\sum_{j=i}^{m} f(z_i;\hat{\mu}_j)}. \tag{11}$$

The connection with the ODP rule is apparent by computing the posterior odds, $v_i/(1-v_i) \approx \Sigma_{\hat{\mu}_j\in A^c} f(z_i;\hat{\mu}_j)/\Sigma_{\hat{\mu}_j\in A} f(z_i;\hat{\mu}_j)$. Finally, thresholding $v_i/(1-v_i)$ is equivalent to thresholding

$$\frac{v_i}{1-v_i} + 1 \approx \frac{\sum_{j=1}^{m} f(z_i;\hat{\mu}_j)}{\sum_{\hat{\mu}_j\in A} f(z_i;\hat{\mu}_j)}.$$

This is expression (3) with $w_j = I(\hat{\mu}_j \in A)$.

Recognizing the ODP as an approximate Bayes rule opens two important directions of generalization. First, we can sharpen the approximation to define a slightly improved procedure, at no cost beyond moderate computational effort. We shall do this in Sections 4.2 and 4.3. Second, we can improve the ODP by making it more relevant to the decision problem at hand by modifying features of the underlying loss function; we shall do this in Section 6.

4.2. The Bayesian discovery procedure statistic

We can improve the approximation in expression (11) by conditioning on a cluster configuration s and using cluster-specific point estimates $\hat{\mu}_j^*$. Given model (8)–(9), we can approximate the posterior probability v_i by

$$S_{\mathrm{BDP}}(z_i) = \sum_{j=1}^{m} f(z_i;\hat{\mu}_j) \Big/ \sum_{\hat{\mu}_j\in A} f(z_i;\hat{\mu}_j). \tag{12}$$

The $\hat{\mu}_j$ are point estimates based on expression (9). For example, we could use the posterior means $\hat{\mu}_j = E(\mu_j|z)$. Short of posterior means, we propose to use a partition $s^{(b)}$ to evaluate cluster-specific point estimates $\hat{\mu}_i = \hat{\mu}_{s_i}^*$, using maximum likelihood estimation within each cluster.

The choice of the specific partition $s^{(b)}$ is not critical. Finally, we report test i significant if $S_{\mathrm{BDP}}(z_i) > t$, for some threshold t. Substituting $\hat{\mu}^*_{s_i}$ in equation (12) the S_{BDP} can be interpreted as a multiple-shrinkage version of the S_{ODP}-statistic. For later reference we note that thresholding S_{BDP} is equivalent to thresholding

$$\hat{v}_i \equiv \sum_{\hat{\mu}^*_j \in A^{\mathrm{c}}} \frac{m_j}{m} f(z_i; \hat{\mu}^*_j) \Big/ \sum_j \frac{m_j}{m} f(z_i; \hat{\mu}^*_j). \tag{13}$$

By the earlier argument $\hat{v}_i \approx v_i$.

The nature of approximation (12) is further clarified and formalized by the following result, which justifies the replacement of μ_i by the cluster-specific MLEs $\hat{\mu}^*_j$. Proving asymptotic results of this kind for the DP is generally not easy. We must establish that the posterior mass for the random G is concentrated on a set of random probability measures with a small number of discontinuities relative to m. See, for example, Arratia *et al.* (2003) and Coram and Lalley (2006) for recent discussions. In particular, Coram and Lalley (2006) addressed the problem from the perspective of 'overfitting' for NP Bayes procedures. They conjectured that, in a variety of problems, the critical determinants of the consistency of Bayes procedures are the rate functions in associated large deviations problems. We avoid the technical difficulties by proving the result for a finite dimensional Dirichlet prior, i.e. a random probability measure G_k such that

$$G_k(A) = \sum_{j=1}^{k} p_j \delta_{\mu^0_j}(\cdot), \tag{14}$$

with $(p_1, \ldots, p_k) \sim D(\alpha/k, \ldots, \alpha/k)$, and $\mu^0_j \sim G^*$ as in expression (9). Inference under the DP prior and equation (14) is comparable, since any integrable functional of the DP can be approximated by corresponding functionals of G_k for sufficiently large k (see Ishwaran and Zarepour (2002)). Also Rodriguez *et al.* (2009) discuss the approximation of inference under a DP prior by results under G_k.

Theorem 1. Assume $z_i|\mu_i \sim^{\mathrm{ind}} f(z_i; \mu_i)$, $i = 1, \ldots, m$, and a random-effects distribution $p(\mu_i|G_k) = G_k$ as in expression (9), with G_k defined in equation (14). Assume that f, $h_A(\cdot)$ and $h_{A^{\mathrm{c}}}(\cdot)$ satisfy the conditions for the Laplace approximation for an open set A (see Schervish (1995), section 7.4.3). Then

$$\lim_{m\to\infty} \{p(r_i = 1|z_1, \ldots, z_m)\} = \frac{E\left\{ \sum_{j:\hat{\mu}^*_j \in A^{\mathrm{c}}} (m_j/m)\, f(z_i; \hat{\mu}^*_j)|z \right\}}{E\left\{ \sum_j (m_j/m)\, f(z_i; \hat{\mu}^*_j)|z \right\}}.$$

The expectation is with respect to the posterior distribution over all possible partitions of $\{1, \ldots, m\}$ with at most k clusters and $\hat{\mu}^*_j$ is the cluster-specific MLE.

For a proof, see Appendix A.

4.3. Multigroup comparisons

The definition of the BDP can easily be extended to the general k samples comparison problem. We assume that data for experimental units i, $i = 1, \ldots, m$, are arranged by k distinct groups, and we wish to test whether the k groups share a common sampling model. Let $x_i = \{x_{i1}, \ldots, x_{in}\}$ be a vector of measurements across k experimental conditions, $i = 1, \ldots, m$. We denote the subset of data from each condition by x^l_i, $l = 1, \ldots, k$, $i = 1, \ldots, m$. Alternatively, data in each group may be reduced to statistics $z^l_i \sim f(z^l_i; \mu^l_i)$, and we can write $z = \{z_1, \ldots, z_m\}$, $z_i = \{z^1_i, \ldots, z^k_i\}$,

with similar notation for μ and μ_i. For notational simplicity, we proceed with the case $k=2$. But the arguments hold for general k. The competing hypotheses are $H_0:(\mu_i^1,\mu_i^2)\in A$ against $H_1:(\mu_i^1,\mu_i^2)\notin A$. Typically $A=\{\mu_i^1=\mu_i^2\}$. Under the loss (5) and the NP model

$$z_i^l|\mu_i^l \overset{\text{ind}}{\sim} f(z_i^l;\mu_i^l), \qquad l=1,2, \quad i=1,\ldots,m,$$
$$\mu_i=\{\mu_i^1,\mu_i^2\}|G \overset{\text{IID}}{\sim} G, \qquad \text{with } G\sim \text{DP}(\alpha,G_0),$$

we can proceed as in Section 2.3 and approximate the posterior odds for the ith comparison by

$$\frac{v_i}{1-v_i}\approx\frac{\displaystyle\int_{A^c} f(z;\,\mu)\,\mathrm{d}F_m(\mu)}{\displaystyle\int_{A} f(z;\,\mu)\,\mathrm{d}F_m(\mu)}=\frac{\displaystyle\sum_{i:(\hat\mu_i^1,\hat\mu_i^2)\in A^c} f(z_i^1|\hat\mu_i^1)\,f(z_i^2|\hat\mu_i^2)}{\displaystyle\sum_{i:(\hat\mu_i^1,\hat\mu_i^2)\in A} f(z;\hat\mu_i)}, \tag{15}$$

where $\hat\mu_i^1$, $\hat\mu_i^2$ and $\hat\mu_i$ are appropriate estimates of the relevant parameters within and across conditions. Expression (15) is an estimated ODP statistic for the multicomparison problem, as discussed in Storey *et al.* (2007). As before, substituting cluster-specific estimates $\hat\mu_{s_i}^*$ for a selected partition s defines the corresponding BDP rule.

5. Comparison of S_{ODP} *versus* S_{BDP}

5.1. Simulation study

We conduct a simulation study to compare the ODP with the NP Bayesian approximation that was outlined in the previous sections. We assume $m=2000$ tests of $H_0:\mu_i=0$ *versus* $H_1:\mu\neq 0$ based on a single observation $z_i\sim N(\mu_i,1)$ for each test. The simulation truth is such that half the observations are drawn from the null hypothesis, whereas the other half are sampled from the following probability distribution:

μ_i	-4	-3	-2	-1.5	1.2	2	3	4.5	5.8
p_i	0.02	0.08	0.01	0.01	0.27	0.02	0.08	0.005	0.005

The distribution mimicks the observed distribution of the t-scores in the microarray data example that is considered in Section 5.2, as shown in Fig. 1. We use $A=\{0\}$, and $h_{A^c}(\cdot)=N(0,1)$. We simulated 1000 data sets. For each simulated data set we ran 2000 iterations of a posterior Markov chain Monte Carlo (MCMC) sampler (with 1000 iterations burn-in). The results confirm the obervation in Storey *et al.* (2007) that the S_{ODP} outperforms the uniformly most powerful unbiased procedure in all cases where the alternative means are not arranged in a perfectly symmetric fashion around zero. The S_{BDP} further improves on the S_{ODP} by borrowing strength across comparisons with the multiple shrinkage that is induced by the DP clustering. Fig. 2 shows that, for any threshold of expected FP, the expected TP is comparable and slightly better under the S_{BDP} than under the S_{ODP}. Expectations are over repeated simulations, i.e. the comparison is by the criterion that is being optimized by the oracle version (1) of the ODP. The curves for the S_{BDP} are computed with a random (last) configuration; the true BDP curve refers to the BDP statistics that are computed on the basis of the (known) true configuration. The differences in Fig. 2 are small. However, for many applications with massive multiple comparisons the number of comparisons is much larger, leading to correspondingly larger differences in true positive results. Most importantly, the improvements come at no additional experimental cost.

We also considered a similar comparison by using equation (12) with posterior means substituted for $\hat\mu_i$, and using $A=(-\varepsilon,\varepsilon)$, for several (small) values of ε, instead of the point null

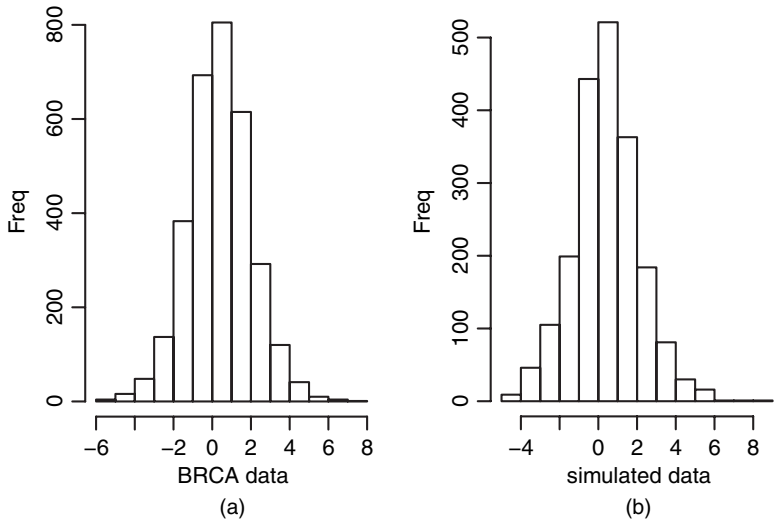

Fig. 1. (a) Scores z_i for the simulation example in Section 5.1 and (b) the data for the BRCA mutation microarray data in Section 5.2

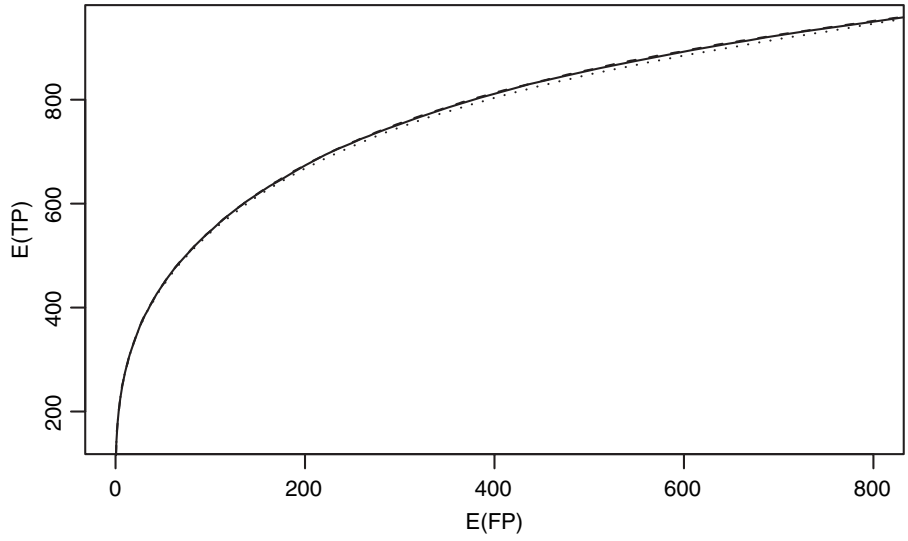

Fig. 2. E(TP) *versus* E(FP) for $\hat{S}_{\mathrm{ODP}}$ and S_{BDP}: ———, BDP(r) (i.e. S_{BDP} computed for a random configuration (we use the last configuration in the MCMC simulation)); - - - - - - -, true BDP (i.e. BDP statistics computed on the basis of the simulation truth); · · · · · · ·, estimated ODP

hypothesis. The plot (which is not shown) of expected TP *versus* FP showed no substantial differences from Fig. 2.

5.2. Microarray data example

We first compare the $\hat{S}_{\mathrm{ODP}}$- *versus* the S_{BDP}-test by analysing a microarray data set that was obtained from breast cancer tumour tissues. The data have been analysed, among others, in Hedenfalk *et al.* (2001), Storey and Tibshirani (2003) and Storey *et al.* (2007) and can be downloaded from `http://research.nhgri.nih.gov/microarray/NEJM_Supplement/`.

The programs that were used to analyse them can be obtained from `http://www.math.unm.edu/~michele/research.html`. The data consist of 3226 gene expression measurements on $n_1 = 7$ BRCA1 breast cancer arrays and $n_2 = 8$ BRCA2 arrays (a third 'sporadic' group was not used for this analysis). Following Storey and Tibshirani (2003), genes with one or more measurement exceeding 20 were eliminated from the data set, leaving $m = 3169$ genes.

Let x_{ij} be the base 2 logarithm expression measurement for gene i on array $j, i = 1, \ldots, m$, $j = 1, \ldots, n$. For illustration, we test differential expression between BRCA1 and BRCA2 mutation genes by using the two-samples t-statistics $z_i = (\bar{x}_{i1} - \bar{x}_{i2})/\sqrt{(s_{i1}^2/n_1 + s_{i2}^2/n_2)}$, where $\bar{x}_{i,k}$ and s_{ik}^2 are respectively the sample mean and sample variance for gene i with respect to the arrays in group k, $k = 1, 2$. We assume model (8)–(9) with $f(z_i; \mu_i) = N(\mu_i, \sigma)$ and test $H_{0i} : \mu_i = 0$. For simplicity, we fix $\sigma = 1$. A model extension with a prior on unknown σ is straightforward. The results remain almost unchanged (they are not shown).

We assess the relative performance of the estimated $\hat{S}_{\mathrm{ODP}}$ against approximation (12) of the NP Bayes rule. For a fair comparison, we evaluate S_{BDP} as a frequentist rule with rejection region $\{S_{\mathrm{BDP}} \geqslant c\}$. The power of the test is evaluated in terms of the FDR and the q-value. See Storey (2002) and Storey *et al.* (2007) for more discussion of the q-value and its evaluation.

The evaluation of S_{BDP} is based on 2000 iterations of a posterior MCMC sampler (1000 burn-in). The number of significant genes that are detected at each q-value is shown in Fig. 3. We report the S_{BDP} as computed on the basis of the cluster configuration of a single iteration of the MCMC sampling scheme. We use the partition from a random iteration and from the maximum *a posteriori* partition. Other choices are possible. However, our experience does not suggest significantly different conclusions by using alternative choices. In Fig. 3 we see that in both cases the S_{BDP} achieves larger numbers of significant genes at the same q-value than the S_{ODP}. The result leads us to recommend the S_{BDP} as a useful tool in multicomparison problems where a natural clustering of the tests is expected. In Table 1, we report the percentage of genes that are flagged by both tests for some choices of q-values. For most q-values of practical interest the S_{BDP}-procedure identifies all genes that were flagged by the S_{ODP}, plus additional discoveries. For example, at $q = 0.05$, the BDP reveals 98 significant genes, against 87 that were revealed by the ODP and 47 by the standard FDR procedure that was devised by Benjamini

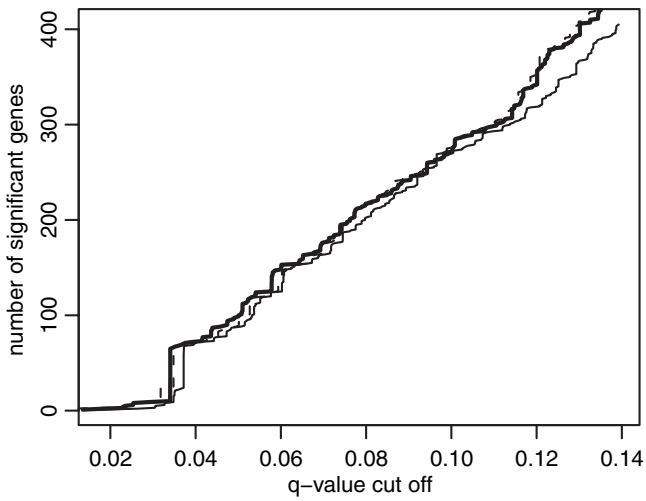

Fig. 3. Comparison of S_{BDP} and S_{ODP} for identifying differentially expressed genes: ———, ODP; - - - - - - -, BDP(r) (i.e. S_{BDP} computed under the (random) last configuration imputed in the MCMC simulation); ———, BDP(h) (i.e. based on the maximum *a posteriori* configuration)

Table 1. Intersection between the decisions with the ODP and the BDP procedures

q-value	*ODP*	*BDP*	*BDP* $\bigcap$ *ODP* (%)
0.05	87	98	100
0.06	124	148	100
0.07	163	176	100
0.08	202	217	100
0.09	234	241	99.5
0.10	272	270	98.14
0.11	293	298	98.63
0.12	318	341	98.11

and Hochberg (1995). Out of the 11 additional genes, seven had been previously reported in the literature as significant in distinguishing BRCA1 and BRCA2 mutations (see Hedenfalk *et al.* (2001)). The additionally identified genes come at no extra cost beyond moderate computational effort. No additional experimental effort is required, and no trade-off in error rates is involved.

For an alternative comparison we consider the 'golden spike' data of Choe *et al.* (2005). They described an Affymetrix GeneChip experiment, where 1309 individual complementary ribonucleic acids have been 'spiked in' at known relative concentrations between the two samples (the spike-in and control samples). We implemented the BDP and ODP as described above. For comparison, we also include alternative comparisons using the significant analysis of microarrays (SAM) procedure (Tusher *et al.*, 2001; Storey, 2002) and independent two-sample *t*-tests. We used the reduced golden spike data set that is provided in the R package `st`. The data set reports 11475 genes with three replicates per group, including 1331 genes with known differential expression. Fig. 4 shows the results of a comparison of BDP, ODP, SAM and independent *t*-tests. To compute the SAM statistic we used the `samr` package in R and alternatively the SAM software from `http://www-stat.stanford.edu/~tibs/SAM/`. Both implementations gave virtually identical results. We caution against overinterpreting the comparison with SAM

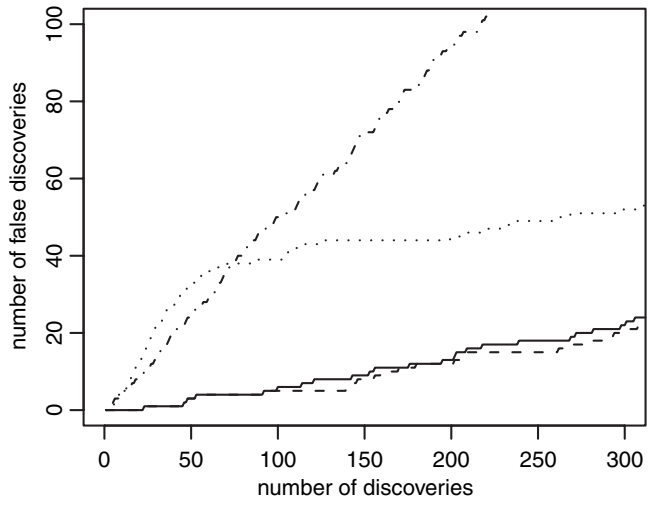

Fig. 4. Golden spike data: comparison of S_{BDP} (———), S_{ODP} (-------), the SAM procedure (·····) and independent two-sample *t*-tests (· – · – ·)

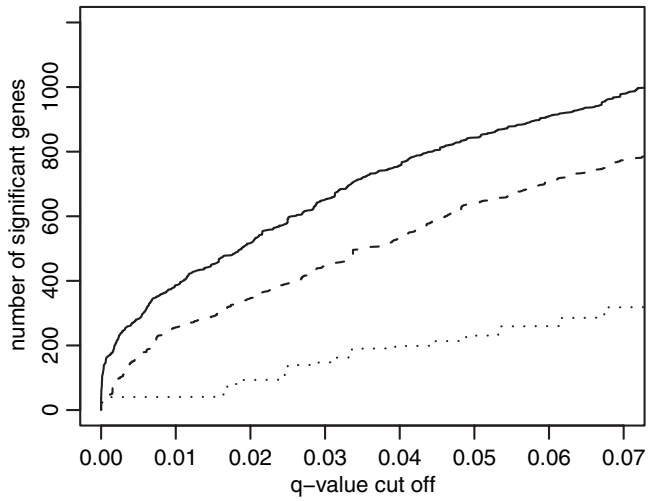

Fig. 5. Hedenfalk data: comparison of S_{BDP} (———), S_{ODP} (-------) and the SAM procedure (·····) under model (16)

in the context of one specific example only. We refer to Storey *et al.* (2007) for more comparative discussion of the ODP *versus* SAM.

Both comparisons, in Figs 3 and 4, are for the ODP and BDP test based on data with a single gene-specific summary statistic z_i per gene. The restriction to a single summary statistic in the definition of the ODP and BDP was purely for presentation and can easily be relaxed. Let x_{ij_k} denote the expression measurement for gene i, sample j and condition k. For simplicity, we still assume only two conditions. Group $k=0$ is assumed to be a control or reference group. Let $\hat{\mu}_{i0}=\Sigma_{l=1}^{n_0} x_{il}/n_0$ be the gene-specific mean of the observations in group 0. We define $z_{ij_k}=x_{ij_k}-\hat{\mu}_{i0}$, $k=0, 1,$ and assume that $z_{ij_k}|\mu_{ik},\sigma_i^2 \sim N(\mu_{ik},\sigma_i^2)$, with $\mu_{i0}\equiv 0$. The hypothesis testing problem becomes

$$H_{0i}:\mu_{i1}=0 \quad \textit{versus} \quad H_{1i}:\mu_{i1}\neq 0, \qquad i=1,\ldots,m.$$

Similarly to before we define a random-effects distribution G for $\theta_i=(\mu_{i1},\sigma_i^2)$ and assume an NP Bayesian prior on G,

$$\theta_i|G\sim G \qquad G\sim \text{DP}(\alpha,G_0), \quad i=1,\ldots,m. \tag{16}$$

For the DP base measure we use a conjugate normal–inverse χ^2-distribution

$$G_0=\{\pi_0\delta_0+(1-\pi_0)N(0,\sigma_i^2/k_0)\}\times \text{Inv-}\chi^2(\sigma_i^2;\nu_0,\sigma_0^{2*}).$$

We can then proceed as in Section 4.3 to define the ODP and BDP statistics. Fig. 5 summarizes the results.

6. Extensions of the optimal discovery procedure and Bayesian discovery procedure

6.1. Weighted loss functions

Once the optimality criterion and the probability model that lead to the ODP have been identified, it is easy to modify the procedure to accommodate preferences and losses other than loss (5). Often some discoveries might be considered more important than others. For example if

$A = \{\mu_i = 0\}$ we might be more interested in large deviations from 0. In this case the loss function should include an explicit reward for detecting true signals as a function of some (monotone) function of μ_i. Scott and Berger (2003) described a decision theoretic approach based on separate loss functions for false positive and false negative results. Similarly, we consider the loss function

$$L(d, \theta, z) = -\sum d_i\, t(\mu_i) + \lambda \sum d_i (1 - r_i), \tag{17}$$

where $t(\mu_i)$ is a known function of the mean, e.g. $t(\mu_i) = \|\mu_i\|$, $\|\cdot\|$ being some norm of μ_i. Let $\overline{t(\mu_i)} = E\{t(\mu_i)|z\}$. The posterior expected loss is

$$E(L|z) = -\sum d_i \{\overline{t(\mu_i)} + \lambda v_i\} + \sum d_i.$$

The Bayes rule is easily shown to be

$$d_i^{m*} = I\{\lambda v_i + \overline{t(\mu_i)} > t\}, \tag{18}$$

for some threshold t. Although the nature of the rule has changed, we can still proceed as before and define a modified S_{ODP}-statistic that approximates the Bayes rule. Let

$$\overline{t(\mu_i)} \approx \sum_{j=1}^{m} t(\hat{\mu}_j) f(z_i; \hat{\mu}_j) \Big/ \sum_{j=1}^{m} f(z_i; \hat{\mu}_j)$$

be an empirical Bayes estimate of $\overline{t(\mu_i)}$, justified similarly to the approximation in expression (11). As before the point estimates $\hat{\mu}_i$ are cluster-specific MLEs $\hat{\mu}^*_{s_i}$ for some partition s. By an argument similar to that before we can justify the following single thresholding statistic as an approximation of rule (18):

$$S^m_{\mathrm{BDP}}(z_i) = \frac{\lambda \sum\limits_{\hat{\mu}_j \notin A} f(z_i; \hat{\mu}_j) + \sum\limits_j t(\hat{\mu}_j)\, f(z_i; \hat{\mu}_j)}{\sum\limits_j f(z_i; \hat{\mu}_j)}. \tag{19}$$

We use $S^m_{\mathrm{BDP}}(y_i)$ as a single thresholding function for the multiple-comparison test in lieu of S_{ODP}.

Any loss function that is written as a sum of comparison-specific terms leads to similar approximations and modifications of the ODP. For example, consider a loss function that involves a stylized description of a follow-up experiment. The loss function is motivated by the following scenario. Many microarray group comparison experiments are carried out as a screening experiment. Genes that are flagged in the microarray experiment are chosen for a follow-up experiment to verify the possible discovery with an alternative experimental platform. For example, Abruzzo *et al.* (2005) described a set-up where reverse transcription polymerase chain reaction experiments are carried out to confirm discoveries proposed by an initial microarray group comparison. Abruzzo *et al.* (2005) reported specific values for correlations across the platforms, error variances etc. On the basis of this set-up Müller *et al.* (2007) considered a loss function that formalizes the consequences of this follow-up experiment. The loss function includes a sampling cost for the reverse transcription polymerase chain reaction experiment and a reward that is realized if the reverse transcription polymerase chain reaction experiment concludes with a significant outcome. The sample size is determined by a traditional power argument for a two-sample comparison, assuming a simple z-test for the difference of two population means. The probability of a significant outcome is the posterior predictive probability of the test statistic in the follow-up experiment falling in the rejection region. Let $(\bar{\mu}_i, s_i)$ denote the posterior mean and standard deviation of the difference in mean expression for gene

i between the two experimental groups. Let $\bar{\rho}$, ρ^* and p_ρ denote known parameters of the joint distribution of the microarray gene expression and the outcome of the reverse transcription polymerase chain reaction experiment for the same gene. Details of the sampling model (see Müller *et al.* (2007)) are not required for the following argument. Finally, let q_α define the $(1-\alpha)$-quantile of a standard normal distribution. For a given signficance level α and a desired power $1-\beta$ at the alternative $\mu_i^A = \bar{\rho}(\bar{\mu}_i - s_i)$, we find a minimum sample size for the follow-up experiment

$$n_i(z_i) = 2\{(q_\alpha + q_\beta)/\mu_i^{\mathrm{A}}\}^2.$$

Let $\Phi(z)$ denote the standard normal cumulative distribution function. The posterior predictive probability for a significant outcome of the follow-up experiment is approximately

$$\pi_i(z_i) = (1-p_\rho)\alpha + p_\rho \Phi\left\{\frac{\rho^* \bar{\mu}_i \sqrt{(n_i/2)} - q_\alpha}{\sqrt{(1+n_i/2\rho^{*2}s_i^2)}}\right\}.$$

We formalize the goal of maximizing the probability of success for the follow-up experiment while controlling the sampling cost by the loss function

$$L(d,\theta,z) = \sum_{d_i=1} \{-c_1\,\pi_i(z_i) + n_i(z_i)\} + c_2 \sum d_i.$$

Here c_2 is a fixed cost per gene for setting up a follow-up experiment, c_1 is the (large) reward for a significant outcome in the follow-up experiment and $c_3 \equiv 1$ is the sampling cost per gene and experiment. The Bayes rule is $d_i^{p*} = I(c_1\pi_i - n_i \geqslant c_2)$. As before we can use

$$\bar{\mu}_i \approx \sum_{j=1}^{m} \hat{\mu}_j f(z_i;\hat{\mu}_j) \Big/ \sum_{j=1}^{m} f(z_i;\hat{\mu}_j)$$

and

$$s_i^2 \approx \sum_{j=1}^{m} (\bar{\mu}_i - \hat{\mu}_j)^2 f(z_i|\hat{\mu}_j) \Big/ \sum_{j=1}^{m} f(z_i;\hat{\mu}_j)$$

and approximate the Bayes rule by a modified ODP style statistic. Let $\hat{\pi}_i$ and $\hat{n}_i$ denote π_i and n_i evaluated with the approximations for $\bar{\mu}_i$ and s_i^2. We consider the modified ODP threshold statistic

$$S^p_{\mathrm{BDP}}(z_i) = c_1\hat{\pi}_i(z_i) - \hat{n}_i.$$

Fig. 6 compares the exact Bayes rules (18) and $d_i^{p*} = I(c_1\pi_i - n_i \geqslant c_2)$ with the tests based on the approximate ODP statistic S^m_{BDP} and S^p_{BDP} respectively.

6.2. Spatial dependence

The nature of the ODP as an approximate Bayes rule was based on the semiparametric model (7). However, the Bayes rules (6) or (18) remain meaningful under any probability model, as long as v_i and $\overline{t(\mu_i)}$ have meaningful interpretations. For example, in geostatistical applications, we may be interested in isolating the exceedance regions of a spatial process, i.e. where the process has values above a given threshold (Zhang *et al.*, 2008). Similarly, in the analysis of functional MRI data, we aim to detect region-specific activations of the brain. See Pacifico *et al.* (2004) and Flandin and Penny (2007) for two recent Bayesian solutions to the problem. In particular, Friston and Penny (2003) proposed an empirical Bayes approach to build posterior probability maps of site-specific signals. These approaches do not make use of an explicitely defined optimality criterion to support the proposed decision rules.

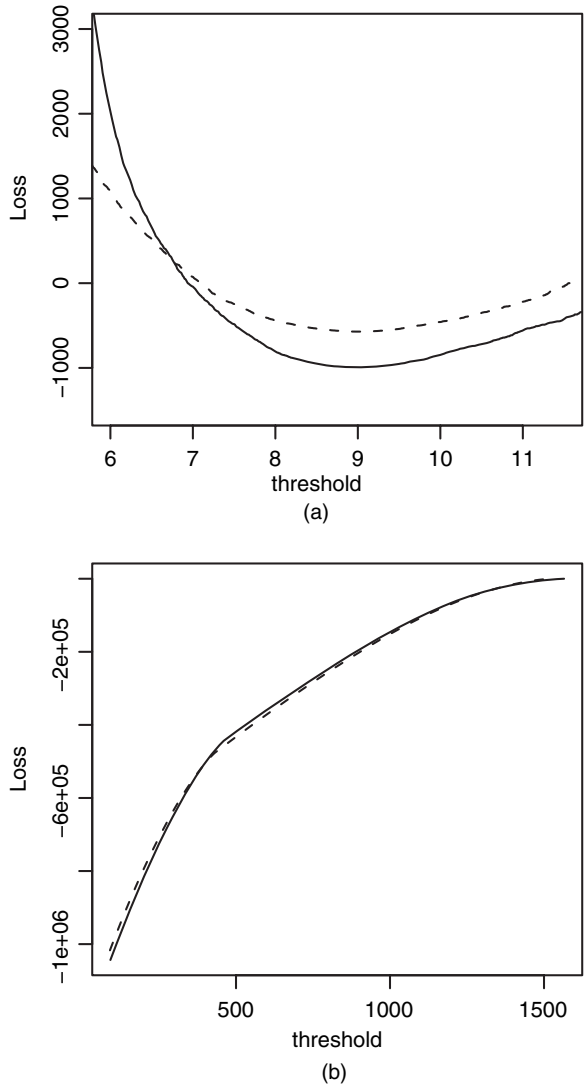

Fig. 6. Expected loss under the exact Bayes rules (———) and expected loss under the ODP rules (-------) plotted against cut-off t: (a) rules d^{m*} and S^m_{BDP}; (b) rules d^{p*} and S^p_{BDP}

We consider a variation of the ODP that is suitable for spatial inference problems, using a specific spatial probability model as an example. We use the spatial model that was proposed by Gelfand *et al.* (2005). Let $\{Y(s), s \in D\}$ be a random field, where $D \subset R^d$, $d \geqslant 2$. Let $s^{(n)} = (s_1, \ldots, s_n)$ be the specific distinct locations in D where observations are collected. Assume that we have replicate observations at each location so that the full data set consists of the collection of vectors $Y_i = (Y_i(s_1), \ldots, Y_i(s_n))^{\mathrm{T}}$, $i = 1, \ldots, m$. We assume that

$$Y_i|\theta_i \overset{\text{ind}}{\sim} f(y_i|\mu + \theta_i), \qquad i = 1, \ldots, m, \tag{20}$$

where f is some multivariate distribution, μ is a (not necessarily constant across s) regressive term and $\theta_i = (\theta_i(s_1), \ldots, \theta_i(s_n))^{\mathrm{T}}$ is a spatial random effect, such that

$$\theta_i | G \overset{\text{IID}}{\sim} G, \qquad i = 1, \ldots, m, \text{ with } G \sim \text{DP}(\alpha, G_0), \tag{21}$$

for some base measure G_0. See Gelfand *et al.* (2005) for details. The assumption of the DP prior in the model is unrelated to the DP that we used to justify the nature of the ODP as the approximate Bayes rule. In this setting, the inferential problem might be quite general, as it may involve subsets of sites and replicates.

For simplicity, we consider a null hypothesis that is specific to each location s and replicate i: $H_{0si} : \theta_i \in A_{si}$, $A_{si} = \{\theta_i(s) > b\}$. For a fixed replicate i, let $d_j = d(s_j)$ be the decision at site s_j, $j = 1, \ldots, n$. Analogously, let $r_j = r(s_j)$ denote the unknown truth at s_j. We could now proceed as before and consider loss (5) and rule $d_j^* = I(v_j > t)$, where v_j is the posterior probability of the event A under the chosen probability model. For m sufficiently large and under model (20)–(21), it is possible to use the asymptotic arguments that were detailed in Section 4 and to define a BDP statistic for the spatial testing problem.

Loss function (5) is usually not an adequate representation of the investigator's preferences in a spatial setting. Posterior probability maps may show very irregular patterns that could lead to, for example, flagging very irregular sets of pixels for exceedance $\theta_i(s) > b$. We may explicitly include in the loss function a penalty for such irregularities, i.e.

$$L(d, \boldsymbol{\theta}, \mathbf{y}) = -\sum d_j r_j + \lambda \sum d_j (1 - r_j) + \gamma \, \text{PI}, \tag{22}$$

where PI is a penalization for irregularities. For example, PI could be the number of connected regions. See the example below. The decision rule corresponding to function (22) is

$$d^*(D) = \underset{\{d(s), s \in D\}}{\arg\min} \, [L\{d(s), \theta(s), y(s)\}].$$

Finding d^* requires numerical optimization.

We illustrate the approach with a data set of 18 individuals who underwent an MRI scan to detect signals of neurodegenerative patterns that are typical of Alzheimer's disease (Ashburner *et al.*, 2003). The data have been provided by the Laboratory of Epidemiology and Neuroimaging, Istituto di Ricovero e Cura a Carattere Scientifico, Centro San Giovanni di Dio, Brescia, Italy, and have been previously normalized with the freely available SPM5 software (`http://www.fil.ion.ucl.ac.uk/spm/`; see Friston *et al.* (1995) and Worsley and Friston (1995)). For simplicity, the data set is restricted to grey density matter intensity values that were collected on a regular two-dimensional grid of 14×19 pixels encompassing the hippocampus and are treated as continuous. The data have been analysed in Petrone *et al.* (2009) before, although with a different inferential aim.

We assume the random-effect model (20)–(21), where f is a multivariate Gaussian distribution, with mean $\mu + \theta$ and covariance matrix $\tau^2 I_n$. The base measure G_0 is a zero-mean stationary Gaussian process with variance σ^2 and correlation $\rho(s, s') = \exp(-\phi \|s - s'\|)$, for some range parameter ϕ. Vague inverse gamma prior distributions on τ^2 and σ^2, and a vague gamma prior for ϕ complete the model. Hence, model (20)–(21) defines a DP mixture of spatial processes (Gelfand *et al.*, 2005). The model is sufficiently flexible to account for most of the spatial dependence that is observed in each individual. However, it is known that one of the marks of Alzheimer's disease is local hippocampal atrophy. Low grey matter intensity observed in normal neuroanatomical structures of the brain should not be reported as a signal. This consideration may lead to the introduction of several kinds of penalties into function (22) to penalize for detections in non-interesting regions. Local atrophy is a condition that typically affects clusters of sites at the same time. This leads us to consider a penalty PI for the number of isolated signals on D. Specifically, PI is the number of interconnected regions and isolated

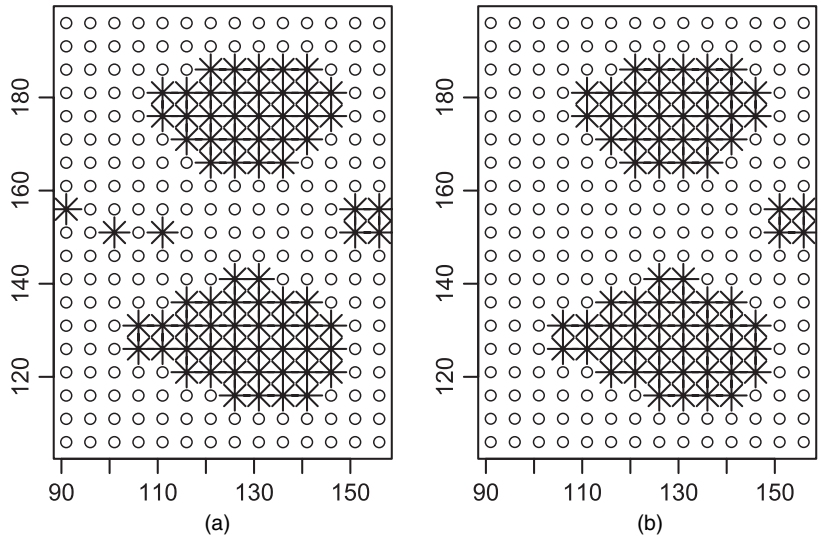

Fig. 7. Effect of a loss function disfavouring isolated signals on the decisions that are taken according to loss (5): (a) optimal decision d^* under $\gamma = 0$; (b) optimal decision d^* with $\gamma = \frac{1}{2}\lambda$

points for which $d_i(s) = 1$. We use a numerical procedure to explore the action space and to minimize equation (22). We find the optimal decision d^* by a random search, initialized with the optimal rule under $\gamma = 0$.

Fig. 7 shows the resulting optimal rule for one individual in the MRI data set. We are interested in detecting regions of low grey matter intensity in the MRI scans. Hence we consider $A = \{\theta(s) < b\}$, where b is a fixed constant, corresponding to the first decile of the data set. The activation threshold for the posterior probability is $t = 0.8$. Fig. 7 shows the activation map for an individual with recognizable signs of hippocampal atrophy for $\gamma = 0$ (Fig. 7(a)) and for $\gamma = \frac{1}{2}\lambda$ (Fig. 7(b)). The additional penalty term provided a principled and coherent means of removing the singleton clusters that would otherwise be reported.

7. Conclusions and summary

Starting from a decision theoretic framework, we provided an interpretation of the ODP as an approximate Bayes rule and introduced two directions of generalizations. First we improved the rule by sharpening the approximation. In a simulation example and a data analysis example we showed improved performance of the resulting BDP rule, even by the frequentist operating characteristics that are optimized by the oracle version of the ODP. Second, we considered generalizations of the ODP by replacing the original generic loss function by loss functions that better reflect the goals of the specific analysis. For loss functions with similar additive structure to that of the original loss function the resulting rule can still be approximated by a single thresholding statistic which is similar to the ODP.

The use of a decision theoretic framework provides a convenient assurance of coherent inference for the approach proposed. However, it also inherits the limitations of any decision theoretic procedure. The optimality is always with respect to an assumed probability model and loss function. The stated loss function is usually a stylized version of the actual goals of inference. Often that is sufficient to obtain a reasonable rule. But we still caution that it is important to validate critically and if necessary to revise the inference in the light of evaluations such as frequentist

operating characteristics. Also, the methods proposed are more computationally intensive than the original ODP procedure. In the simplest case we require some additional simulation to find a clustering of comparisons to compute cluster-specific MLEs.

The main strengths of the approach proposed are the generality and the assurance of coherent inference. The approach is general in the sense that the methods proposed are meaningful for any underlying probability model, and in principle for arbitrary loss functions. The approach is coherent in the sense that it derives from minimizing expected loss under a well-defined probability model. From a data analysis perspective, an important strength of the approach is the need and the opportunity to think explicitly about the inference goals and to formalize them in the loss function. A practical strength is the opportunity for improved inference at no additional experimental cost, and only moderate additional computational cost.

Aknowledgement

The work of the third author is partly supported by National Institutes of Health clinical and translational science awards grant UL1 RR024982.

Appendix A

A.1. Proof of proposition 1

The proof follows closely the proof of theorem 1 in Storey and Tibshirani (2003). First, rewrite

$$\text{pFDR} = E\left(\frac{\text{FP}}{D} | D>0\right) = E\left\{\frac{\sum_{i=1}^{n} d_i(1-r_i)}{\sum_{i=1}^{n} d_i} | D>0\right\}. \tag{23}$$

The expectation is with respect to the distribution of $(z_1, \ldots, z_m)$, conditionally on the event that some of the comparisons are significant. Hence,

$$\text{pFDR} = E_{Z_1,\ldots,Z_m|D>0}\left[E\left\{\frac{\sum_{i=1}^{n} d_i(1-r_i)}{\sum_{i=1}^{n} d_i} | z_1, \ldots, z_m\right\}\right].$$

Since d_i is a function of the sample $z_1, \ldots, z_m$, the inner expectation is just

$$E\left\{\frac{\sum_{i=1}^{n} d_i(1-r_i)}{\sum_{i=1}^{n} d_i} | z_1, \ldots, z_m\right\} = \frac{\sum_{i=1}^{n} d_i(1-v_i)}{\sum_{i=1}^{n} d_i},$$

and, since $d_i^* = I(v_i > t)$,

$$\text{pFDR} < E_{Z_1,\ldots,Z_m|D>0}\left\{\frac{\sum_{i=1}^{n} d_i(1-t)}{\sum_{i=1}^{n} d_i}\right\} = 1-t.$$

A.2. Proof of theorem 1

Because of the exchangeability of samples from a Pólya urn, without loss of generality, we may consider $v_m = p(r_m = 1 | z_1, \ldots, z_m)$. First, note that

$$v_m = \int p(r_m = 1 | G_k, z_1, \ldots, z_m)\, p(\mathrm{d}G_k | z_1, \ldots, z_m) = \int G_k(A^c)\, p(\mathrm{d}G_k | z_1, \ldots, z_m)$$
$$= \frac{\int \int_{A^c} p(z_m | \mu_m)\, G_k(\mathrm{d}\mu_m)\, p(\mathrm{d}G_k | z_1, \ldots, z_{m-1})}{p(z_m | z_1, \ldots, z_{m-1})}.$$

Both numerator and denominator take the form

$$E\left\{ \int_B p(z_m | \mu_m)\, G_k(\mathrm{d}\mu_m) | z_1, \ldots, z_m \right\} = \int_{\mathcal{P}} \int_B p(z_m | \mu_m)\, G_k(\mathrm{d}\mu_m)\, p(\mathrm{d}G_k | z_1, \ldots, z_m),$$

for a Borel set B. Let $s^{0(m)}$ be a vector of cluster indicators, i.e. $s_i^0 = j$ if and only if $\mu_i = \mu_j^0$, $i = 1, \ldots, m$, $j = 1, \ldots, k$.

For any fixed m, let $\{s^{0(m)}\}$ denote the set of all partitions of $\{1, \ldots, m\}$ into at most k clusters. From the discussions in Ferguson (1983), Lo (1984) and Ishwaran and James (2003), it follows that

$$E\left\{ \int_B p(z_m | \mu_m)\, G_k(\mathrm{d}\mu_m) | z_1, \ldots, z_{m-1} \right\} = \frac{m}{\alpha + m} \sum_{s^{(m-1)}} p(s^{(m-1)} | z_1, \ldots, z_{m-1})$$
$$\times \sum_{j=1}^{k} \frac{\alpha/K + m_j^0}{m} \int_B p(z_m | \mu_j^0)\, p(\mathrm{d}\mu_j^0 | z_i : s_i = j, i = 1, \ldots, m-1), \tag{24}$$

where $m_j^0 = \mathrm{card}\{s_i^0 : s_i^0 = j, i = 1, \ldots, m-1\}$ is the number of elements in cluster j. If $m_j^0 = 0$, then $p(\mathrm{d}\mu_j^0 | z_i : s_i = j) \equiv G^*(\mathrm{d}\mu_j^0)$. Expression (24) highlights that any partition of $\{1, \ldots, m\}$ can be obtained from a corresponding partition of $\{1, \ldots, m-1\}$ by adding the mth observation to any of the previous clusters or by forming a new cluster.

Now, note that, for each $j = 1, \ldots, k$, either $m_j^0/m = o(1)$ or $m_j^0/m = O(1)$. If $m_j^0/m = o(1)$, $(\alpha/k + m_j^0)/m \to 0$; if $m_j^0/m = O(1)$, we can use Laplace approximation arguments (see Schervish (1995), section 7.4.3, or Ghosh *et al.* (2006), page 115) to obtain

$$\int_B p(z_m | \mu_j^0)\, p(\mathrm{d}\mu_j^0 | z_i : s_i^0 = j) \approx p(z_m | \hat{\mu}_j^0)\, \Phi(\hat{\mu}_j^0 \in B)\{1 + O(m_j^{0-2})\},$$

where $\hat{\mu}_j^0$ is the MLE computed in cluster j, $j = 1, \ldots, k$, obtained by solving

$$\frac{\partial}{\partial \mu} \sum_{i : s_i = j} f(z_i; \mu) + \frac{\partial}{\partial \mu} f(z_m; \mu) = 0$$

and $\Phi(\cdot)$ denotes a standard Gaussian probability distribution.

Next we relabel the non-empty clusters by identifying the set $\{\mu_j^0; m_j^0 > 0\}$ as the set of unique values $\{\mu_j^*, j = 1, \ldots, L\}$.

The proof is completed after noting that, since $m_j^0/m = O(1)$ and, because of the asymptotic consistency of posterior distributions, as $m \to \infty$, $\Phi(\hat{\mu}_j \in B) \to I_B(\hat{\mu}_j)$.

References

Abruzzo, L. V., Lee, K. Y., Fuller, A., Silverman, A., Keating, M. J., Medeiros, L. J. and Coombes, K. R. (2005) Validation of oligonucleotide microarray data using microfluidic low-density arrays: a new statistical method to normalize real time RT-PCR data. *Biotechniques*, **38**, 785–792.

Arratia, R., Tavaré, S. and Barbour, A. (2003) *Logarithmic Combinatorial Structures: a Probabilistic Approach*. Helsinki: European Mathematical Society.

Ashburner, J., Csernansky, J., Davatzikos, C., Fox, N., Frisoni, G. and Thompson, P. (2003) Computer-assisted imaging to assess brain structure in healthy and diseased brains. *Lancet Neurol.*, **2**, 79–88.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Bogdan, M., Gosh, J. and Tokdar, S. (2008) A comparison of the benjamini-hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (eds N. Balakrishnan, E. Peña and M. Silvapulle), pp. 211–230. Beachwood: Institute of Mathematical Statistics.

Cao, J., Jie, X., Zhang, S., Whitehurst, A. and White, M. (2009) Bayesian optimal discovery procedure for simultaneous significance testing. *BMC Bioinform.*, **10**, no. 5.
Choe, S., Boutros, M., Michelson, A., Church, G. and Halfon, M. (2005) Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biol.*, **6**, article R16.
Cohen, A. and Sackrowitz, H. (2007) More on the inadmissibility of step-up. *J. Multiv. Anal.*, **98**, 481–492.
Coram, M. and Lalley, S. (2006) Consistency of bayes estimators of a binary regression function. *Ann. Statist.*, **34**, 1233–1269.
Dahl, D. and Newton, M. (2007) Multiple hypothesis testing by clustering treatment effects. *J. Am. Statist. Ass.*, **102**, 517–526.
Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.
Ferguson, T. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.
Ferguson, T. (1974) Prior distributions on spaces of probability measures. *Ann. Statist.*, **2**, 615–629.
Ferguson, T. S. (1983) Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* (eds H. Rizvi and J. Rustagi), pp. 287–302. New York: Academic Press.
Flandin, G. and Penny, W. (2007) Bayesian fmri data analysis with sparse spatial basis function priors. *Neuroimage*, **34**, 1108–1125.
Friston, K. J., Ashburner, L., Poline, J., Frith, C. and Frackowiak, R. (1995) Spatial registration and normalization of images. *Hum. Brain Mappng*, **2**, 165–189.
Friston, K. J. and Penny, W. (2003) Posterior probability maps and spms. *Neuroimage*, **19**, 1240–1249.
Gelfand, A., Kottas, A. and MacEachern, S. (2005) Bayesian nonparametric spatial modeling with Dirichlet processes mixing. *J. Am. Statist. Ass.*, **100**, 1021–1035.
Ghosh, J., Delampady, M. and Tapas, S. (2006) *An Introduction to Bayesian Analysis: Theory and Methods*. New York: Springer.
Gopalan, R. and Berry, D. (1993) Bayesian multiple comparisons using dirichlet process priors. *J. Am. Statist. Ass.*, **93**, 1130–1139.
Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, A. and Trent, J. (2001) Gene-expression profiles in hereditary breast cancer. *New Engl. J. Med.*, **344**, 539–549.
Ishwaran, H. and James, L. (2003) Some further developments for stick-breaking priors: finite and infinite clustering and classification. *Sankhya* A, **65**, 577–592.
Ishwaran, H. and Zarepour, M. (2002) Dirichlet prior sieves in finite normal mixtures. *Statist. Sin.*, **12**, 941–963.
Lo, A. (1984) On a class of bayesian nonparametric estimates: I, density estimates. *Ann. Statist.*, **12**, 351–357.
Müller, P., Parmigiani, G. and Rice, K. (2007) Fdr and bayesian multiple comparisons rules. In *Bayesian Statistics 8* (eds J. M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West). Oxford: Oxford University Press.
Müller, P., Parmigiani, G., Robert, C. P. and Rousseau, J. (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Am. Statist. Ass.*, **99**, 990–1001.
Neal, R. M. (2000) Markov chain sampling methods for dirichlet process mixture models. *J. Computnl Graph. Statist.*, **9**, 249–265.
Pacifico, M. P., Genovese, C., Verdinelli, I. and Wasserman, L. (2004) False discovery control for random fields. *J. Am. Statist. Ass.*, **99**, 1002–1014.
Petrone, S., Guindani, M. and Gelfand, A. (2009) Hybrid Dirichlet mixture models for functional data. *J. R. Statist. Soc.* B, **71**, 755–782.
Rodriguez, A., Dunson, D. and Gelfand, A. (2008) The nested Dirichlet process. *J. Am. Statist. Ass.*, **103**, 1131–1154.
Schervish, M. J. (1995) *Theory of Statistics*. New York: Springer.
Scott, J. and Berger, J. (2003) An exploration of aspects of bayesian multiple testing. *J. Statist. Planng Inf.*, **136**, 2144–2162.
Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.
Storey, J. D. (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc.* B, **69**, 347–368.
Storey, J., Dai, J. and Leek, J. (2007) The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, **8**, 414–432.
Storey, J. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natn. Acad. Sci. USA*, **100**, 9440–9445.
Tusher, V., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natn. Acad. Sci. USA*, **98**, 5116–5121.
Worsley, K. J. and Friston, K. J. (1995) Analysis of fmri time-series revisited—again. *Neuroimage*, **2**, 173–181.
Zhang, J., Craigmile, P. and Cressie, N. (2008) Loss function approaches to predict a spatial quantile and its exceedance region. *Technometrics*, **50**, 216–227.