A Bayesian Discovery Procedure

Michele Guindani

University of New Mexico, Alberquerque, NM 87111, U.S.A.

Peter Müller

University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, U.S.A.

Song Zhang

University of Texas Southwestern Medical Center, Dallas, TX 75390, U.S.A.

Summary. The optimal discovery procedure (ODP) maximizes the expected number of true positives for every fixed expected number of false positives. We show that the ODP can be interpreted as an approximate Bayes rule under a semi-parametric model. Improving the approximation leads us to a Bayesian discovery procedure (BDP), which exploits the multiple shrinkage in clusters implied by the assumed nonparametric model. We compare the BDP and the ODP estimates in a simple simulation study and in an assessment of differential gene expression between two tumor samples. We extend the setting of the ODP 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.

Key words: MULTIPLE COMPARISON, FALSE DISCOVERY RATE, LOSS FUNCTION, BAYES OPTIMAL RULE.

1. Introduction.

A number of different 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 (FWER) 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), 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 different outcomes according to the preferences and inferential focus of the investigators. Cohen and Sackrowitz (2007) prove the inadmissibility of the Benjamini and Hochberg procedure under any loss expressed as a linear combination of false discoveries and false acceptances and under several sampling models, including the general one-parameter exponential family. Müller et al. (2007) undertake a decision theoretic approach to multiple testing and discuss several loss functions that lead to the use of FDR-based rules.

The general multiple comparison problem is stated as follows. Assume 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 is 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 we wish to test $\mu_i \in A$ vs. $\mu_i \notin A$ for $i = 1, \ldots, m$. Storey (2007a) proposed the optimal discovery procedure (ODP). The ODP is based on a single significance thresholding statistic,

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

The null hypothesis H_{0i} is rejected if $S_{\mathrm{ODP}}(z_i) \geq \lambda$, for some $0 \leq \lambda < \infty$. We write ${d_i}^{\mathrm{ODP}} = I(S_{\mathrm{ODP}}(z_i) \geq \lambda)$. For a point null, $A = \{0\}$, the test reduces to thresholding $S_{\mathrm{ODP}}(z_i) = \sum_{\mu_j \notin A} f(z_i; \mu_j) / f(z_i; 0)$. Storey proves that $d^{\ \mathrm{ODP}}$ maximizes the expected number of true positives (ETP) among all procedures with equal or smaller expected number of false positives (EFP). The threshold function S_{ODP} involves the unknown parameters μ_j, $j = 1, \ldots, m$. In practice, $S_{\mathrm{ODP}}(\cdot)$ has to be estimated. When every test has the same null distribution and there are no nuisance parameters, i.e., $A = \{0\}$, the ODP is evaluated as

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

where $\widehat{\mu}_j$ is a point estimate for μ_j (e.g., the maximum likelihood estimate). It is shown that the performance of $\hat{S}_{\text{ODP}}$ is comparable to the theoretically optimal discovery procedure based on S_{ODP}.

For general A the ODP proceeds with the following thresholding function,

$$\hat{S}_{\text{ODP}}(z) = \frac{\sum_{j=1}^{m} f(z; \widehat{\mu}_j)}{\sum_{j=1}^{m} w_i f(z; \widehat{\mu}_j)}, \tag{3}$$

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 to be null, and $w_j = 0$ otherwise. Storey et al. (2007b) show that (3) outperforms many procedures commonly used in testing a large number of genes for differential expression (e.g., SAM, empirical Bayes "local" FDR).

The ODP statistic can be immediately recognized as an empirical Bayes factor. However, the nature of the rule and the extent of the approximation are better appreciated by recognizing $d^{\text{ ODP}}$ as an approximate Bayes rule under a nonparametric prior model for μ_i. This result is in accordance with Ferguson's (1973) observation that nonparametric Bayesian inference often yields results that are comparable to corresponding classifical inference. To be more specific, we define 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 (RPM). A prior probability model for an RPM is known as a non-parametric Bayesian model. In particular, we assume a Dirichlet process (DP) prior, one of the most popular nonparametric Bayesian models.

We show that the Bayes rule that maximizes the expected number of true positives subject to a bound on the expected number of false positives is a threshold on the marginal posterior probability $v_i = Pr(\mu_i \notin A \mid data)$. Under the DP prior, for sufficiently large m, this marginal posterior probability can be approximated in a way that justifies the use of $d^{\text{ ODP}}$ as an approximate

Bayes rule. The previous argument also provides a simple alternative proof of the claimed ODP optimality subject to some approximations. The expectation in Storey's optimality statement for $d^{\text{ ODP}}$ 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 positives. A bound on the conditional expectation, conditional on each outcome z, implies the same bound on the marginal expectation. Hence, we conclude that the Bayes rule under the nonparametric prior approximately satisfies the optimality property of the ODP procedure. Besides, by the same arguments, we show that thresholding the marginal posterior probability amounts to controlling the positive FDR (Storey, 2002; Storey and Tibshirani, 2003).

DP priors in the context of multiple hypotheses testing have been considered before by Gopalan and Berry (1993). More recently, Dahl and Newton (2007) have proposed a DP mixture model (BEMMA) for testing correlated hypotheses and showed that the induced clustering information leads to an increased testing power. The distinction between the previous approaches and ours is that here the DP prior is part of a decision theoretic setup. Besides, both Gopalan and Berry (1993) and Dahl and Newton (2007) restrict inference to hypotheses concerning the configuration of ties among the parameters of interest.

Once we have established the utility function and the probability model leading to the optimal discovery procedure, we can proceed with generalizations in two directions. First, we will 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 MRI scans.

In section 2, we introduce the decision problem and discuss the non-parametric

(NP) probability model that leads to the Bayesian interpretation of the ODP and the corresponding BDP statistics. In section 3, we provide details on how to obtain the BDP statistics when we assume a conjugate DP mixture of normal model for the data. In section 4, we compare the behavior of the ODP and the BDP with a simulated and a microarray dataset, and show that the BDP provides at least some improvement on the optimal procedure. In section 5, we discuss some extensions of the previous settings for multigroup experiments, different loss functions and different inferential purposes. In particular, we consider a dependent NP model for the study of spatial phenomena, and apply the Bayesian multicomparison decision rule to a simplified MRI dataset. Finally, in section 6, we provide some conclusions and discuss further directions of research.

2. The ODP as Approximate Bayes Rule

2.1. The Decision Problem

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

$$H_{0i} : \mu_i \in A \text{ vs. } H_{1i} : \mu_i \notin A,$$

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

$$p(\mu_i \mid G) = \pi_0\, G(\mu_i \mid A) + (1 - \pi_0)\, G(\mu_i \mid A^c),$$

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 i-th comparison.

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

We will 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).

Let $d_i \in \{0,1\}$ denote the decision for the i-th hypothesis, with $d_i = 1$ indicating a decision against the i-th null hypothesis, 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 $d_i^\star(z)$ is defined by minimizing L in expectation with respect to the posterior model $p(\theta \mid z)$. Formally,

$$d^\star = \arg\min_d \int L(d, \theta, z)\, p(\theta \mid z)\, \mathrm{d}\theta.$$

Our aim is to characterize the ODP as an approximate Bayes rule. Therefore, keeping in mind the optimality property of the ODP, we consider a loss function that combines the true positive count, $\text{TP} = \sum d_i\, r_i$, and false positive count, $\text{FP} = \sum d_i(1 - r_i)$,

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

The loss (5) can be interpreted as the Lagrangian for maximizing TP subject to a given bound on FP.

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

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

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. *Suppose m hypothesis tests $H_{0i} : \mu_i \in A$ vs $H_{1i} : \mu_i \in A^c$ are performed with the statistics $z_1, \ldots, z_m$, where $z_i \mid \mu_i \sim f(z_i;\, \mu_i)$, independently across i, $i = 1, \ldots, m$, and $p(\mu_i \mid G) = \pi_0\, G(\mu_i \mid A) + (1 - \pi_0)\, G(\mu_i \mid A^c)$, for some distribution G and probabilty $\pi_0 = prob(H_{0i}) = prob(r_i = 0)$. Let the rejection region be determined by the Bayes rule under the loss function (5), i.e. $d_i^* = I(v_i > t)$, where $v_i = p(r_i = 1|z_1, \ldots, z_m)$. Let $FP = \sum_{i=1}^m d_i(1 - r_i)$ and $D = \sum_{i=1}^m d_i$. Then,*

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

PROOF. See appendix.

Rule (6) is different from the one prescribed by the local FDR, which is defined as the posterior probability of the null given that we have observed a certain value z_i of the statistics, 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 local to z_i, while v_i is a global measure of significance.

2.2. A Semiparametric Bayesian Model

We complete the sampling model (4) with a prior model for G that will allow us to identify $d^\star$ as a rule approximately matching the ODP.

Prior probability models for unknown distributions, G in this case, are traditionally known as non-parametric (NP) Bayesian models. One of the most commonly used NP Bayesian priors is the DP model. We write $G \sim \text{DP}(G^\star, \alpha)$ to indicate a DP for a random probability measure G. See Ferguson (1973) and Ferguson (1974) for a definition and important properties of the DP model. The model requires the specification of two parameters, the base measure $G^\star$ and the total mass parameter α. The base measure $G^\star$ is the prior mean, $E(G) = G^\star$. 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 a.s. discrete. This allows us to write G as a mixture of point masses, $G = \sum w_h \delta_{m_h}$. Another important property is the conjugate nature of the DP prior under random sampling. Assume $\mu_i \sim G$, $i = 1, \ldots, m$, are an i.i.d. sample from a random measure G with DP prior, $p(G) = DP(G^\star, \alpha)$. Then, the posterior probability model is $p(G \mid \mu) = DP(G_1^\star, \alpha + m)$, for $G_1^\star \propto \alpha G^\star + m F_m$. Here, $F_m = 1/m \sum \delta_{\mu_i}(\cdot)$ is the empirical distribution of the realized μ_i's.

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

$$\mu_i | G \sim G \quad G \sim \mathrm{DP}(G^\star, \alpha). \tag{7}$$

Model (7) implies that the prior for the null hypothesis $p_0 = G(A)$ is Beta, $p_0 \sim Be\left(\alpha G^\star(A), \alpha[1 - G^\star(A)]\right)$.

2.3. Posterior Inference and the Bayes Rule

The Bayes rule (6) requires only the evaluation of the marginal posterior probabilities $v_i = P(r_i = 1 \mid z)$. Here z generically denotes the observed data. We show that, under a DP prior model, $d^\star \approx d^{\text{ ODP}}$. We start by writing the marginal posterior probability as expectation of conditional posterior probabilities,

$$v_i = E\left[p(r_i = 1 \mid G, z) \mid z\right] = E\left[\frac{\int_{A^c} f(z_i;\ \mu)\, dG(\mu)}{\int_{A \cup A^c} f(z_i;\ \mu)\, dG(\mu)} \mid z\right],$$

and proceed with an approximation of the conditional posterior distribution $p(G \mid z)$. The conjugate nature of the DP prior under random sampling implies

$$E(G \mid \mu, z) = G_1^\star \propto \alpha G^\star + \sum \delta_{\mu_i}.$$

Recall that $F_m \propto \sum \delta_{\widehat{\mu}_i}$ is the empirical distribution of the maximum likelihood estimates $\widehat{\mu}_i$. For large m, small α, and an informative sampling model $f(z_i; \mu)$, $E(G \mid z) \approx F_m$. Further, for large m, the uncertainty of the posterior DP, $p(G \mid \mu) = DP(G^\star, m{+}\alpha)$ is negligible, allowing us to approximate the posterior on the random probability measure G with a degenerate distribution at F_m, $p(G \mid z) \approx \delta_{F_m}(G)$, i.e. $G \approx \frac{1}{m} \sum \delta_{\widehat{\mu}_i}$. Therefore,

$$v_i \approx \frac{\int_{A^c} f(z;\ \mu)\, dF_m(\mu)}{\int f(z;\ \mu)\, dF_m(\mu)} = \frac{\sum_{\widehat{\mu}_j \in A^c} f(z_i;\ \widehat{\mu}_j)}{\sum_{j=i}^{m} f(z_i;\ \widehat{\mu}_j)}. \tag{8}$$

The connection with the ODP rule is apparent by computing the posterior odds,

$$v_i/(1 - v_i) \approx \frac{\sum_{\widehat{\mu}_j \in A^c} f(z_i;\ \widehat{\mu}_j)}{\sum_{\widehat{\mu}_j \in A} f(z_i;\ \widehat{\mu}_j)}.$$

Finally, thresholding $v_i/(1 - v_i)$ is equivalent to thresholding

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

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

3. The Bayesian Discovery Procedure (BDP)

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 will do this in Sections 3.1 and 3.2. 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 will do this in Section 5.

3.1. BDP for Testing Normal Means.

We outline the algorithm for the BDP in the simple case of a mixture of DP normal model. Many practical methods 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 discuss how these methods can be conveniently adapted in the multiple comparison setting to compute the posterior probabilities v_i and the approximate DP based Bayes rule.

Let z_i, $i = 1, \cdots, m$ denote the observed data (or statistics thereof) for test i. We assume

$$z_i \mid \mu_i \overset{ind}{\sim} N(\mu_i, \sigma),\; i = 1, \cdots, m, \tag{9}$$

where $N(\eta, \tau)$ denotes a normal distribution with mean η and variance τ^2. The NP Bayes model (7) is specified as

$$\begin{aligned} \mu_i \mid G &\overset{iid}{\sim} G \\ G &\sim \text{DP}(G^\star, \alpha) \text{ with } G^\star(\cdot) \sim \pi_0 h_A(\cdot) + (1 - \pi_0) h_{A^c}(\cdot), \end{aligned} \tag{10}$$

where $h_A(\cdot)$ and $h_{A^c}(\cdot)$ are distributions with support, respectively, on A and A^c. Equations (9) and (10) define a DP mixture model where $G^\star$ itself is a mixture of two terms.

Algorithms for posterior Monte Carlo simulation in DP process mixture models can easily be modified to adapt to the mixture in $G^\star$. We will 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 will denote it simply by $h(\cdot)$. The a.s. 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 component indicators excluding the i-th one. Also, let L be the number of clusters 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 s-th cluster, and let $\mu^\star_s = \mu_i$, $i \in \Gamma_s$ denote the common value of μ_i for cluster s. Then, the i-th observation falls in one of the existing clusters or forms a new cluster according to the following modified Pólya urn scheme,

$$P(s_i = s \mid s_{-i}, z) \propto \begin{cases} \frac{m_{-i,1}+\pi_0\alpha}{m-1+\alpha} f(z_i \mid \mu^\star_{s_i} = 0) & \text{if } s = 1 \\ \frac{m_{-i,s}}{m-1+\alpha} \int \, f(z_i \mid \mu^\star_s) \, h(\mu^\star_s | z_{-i,s}) \, d\mu^\star_s, & \text{if } 2 \leq s \leq L \\ \frac{(1-\pi_0)\alpha}{m-1+\alpha} \int \, f(z_i \mid \mu^\star_s) \, h(\mu^\star_s) \, d\mu^\star_s, & \text{if } s = L+1 \end{cases}$$

Here, $h(\mu^\star_s \mid z_{-i,s})$ denotes the posterior distribution of $\mu^\star_s$ based on the prior $h(\cdot)$ and all observations z_h, $h \in \Gamma_s/\{i\}$. Note that the posterior of $\mu^\star_1$ given all the observations in cluster 1 is still a point mass $\delta_0(\cdot)$. There is a positive probability for $m_1 = m$, i.e., that all the elements fall in the single cluster $A = \{0\}$. This can lead to a practically absorbing state in the MCMC simuulation. However, from Antoniak (1974), it follows that such a probability tends to zero as m increases. If the problem should arise in practice, multiple chains might be used to overcome the obstacle. We described the algorithm for a particular choice of A and $G^\star$. But it can be easily extended to more general A and $G^\star$. 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 $\pi_0 = G^\star(A)$. Then, equations (9) and (10) describe a customary mixture of DP normal model with Gaussian base measure $G^\star$ and 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 we have B posterior Monte Carlo samples of random partitions, $s^{(b)} = (s^{(b)}_1, \ldots, s^{(b)}_m)$, $b = 1, \ldots, B$. We compute an estimate of $v_i = E(r_i \mid z)$ as $\hat{v}_i = 1 - \sum_{b=1}^{B} I(s^{(b)}_i = 1)/B$, that is the proportion of iterations where test i has not been assigned to the null cluster. Similarly, we can compute the NP Bayesian approximation of S_{ODP} as

$$S_{\text{BDP}}(z_i) = \frac{\sum_{j=1}^{m} f(z_i; \, \widehat{\mu}_j)}{\sum_{\widehat{\mu}_j \in A} f(z_i; \, \widehat{\mu}_j)}. \tag{11}$$

The $\widehat{\mu}_j$ are point estimates based on (10). For example, one could use the posterior means $\widehat{\mu}_j = E(\mu_j \mid z)$. Short of posterior means, we propose to use a partition $s^{(b)}$ to evaluate cluster-specific point estimates $\widehat{\mu}_i = \widehat{\mu}^{\star}_{s_i}$, using maximum likelihood estimation within each cluster. The choice of the specific configuration $s^{(b)}$ is not critical. Finally, we report test i significant if $S_{\text{BDP}}(z_i) > t$, for some threshold t. The choice of t depends on the form of the loss function.

Substituting $\widehat{\mu}^{\star}_{s_i}$ in (11) the S_{BDP} can be interpreted as a multiple shrinkage version of the S_{ODP} statistic.

Thresholding S_{BDP} is equivalent to thresholding

$$\widehat{v}_i \equiv \sum_{\widehat{\mu}^*_j \in A} m_j/m\, f(z_i;\, \widehat{\mu}^{\star}_j) \bigg/ \sum_j m_j/m\, f(z_i;\, \widehat{\mu}^{\star}_j). \tag{12}$$

By the earlier argument $\widehat{v}_i \approx v_i$. The nature of the approximation is further clarified and formalized by the following result. We prove the result for a finite truncation of the Dirichlet Process, that is a random probability measure G_k such that

$$G_k(A) = \sum_{j=1}^{k} p_j \delta_{\mu^o_j}(\cdot), \tag{13}$$

with $p_j = V_j \prod_{l=1}^{j-1}(1 - V_l)$, $V_l \sim Beta(\alpha, 1)$, $l = 1, \ldots, k-1$, $V_k = 1$, and $\mu^o_j \sim G^*$ as in (10). Inference under the two models is comparable, since any integrable functional of the Dirichlet Process can be approximated by corresponding functionals of G_k for sufficiently large k (see Iswharan and Zarepour, 2002). Also Rodriguez et al. (2008) discuss the approximation of inference under a DP prior by results under G_k. Asymptotically, thresholding $\hat{v}_i$ in (12) is equivalent to thresholding the Bayes rule (6), under the NP Bayes model (9-10) and loss function (5).

Theorem 2. *Assume $z_i|\mu_i \overset{ind}{\sim} f(z_i; \mu_i)$, $i = 1, \ldots, m$ and a random effects distribution $p(\mu_i \mid G_k) = G_k$ as in (10), with G_k defined in* (13). *Assume f, $h_A(\cdot)$, and $h_{A^c}(\cdot)$ satisfy the conditions for the Laplace approximation for an open set A (see Schervish, 1995, chapter 7.4.3). Then*

$$\lim_{m \to +\infty} p(r_i = 1|z_1, \ldots, z_m) = \frac{E\left(\sum_{j:\widehat{\mu}^{\star}_j \in A} \frac{m_j}{m} f(z_i;\, \widehat{\mu}^{\star}_j)\right)}{E\left(\sum_j \frac{m_j}{m} f(z_i;\, \widehat{\mu}^{\star}_j)\right)},$$

where the expectation is taken with respect to the posterior distribution over all the possible partitions of $\{1, \ldots, m\}$ *with at most* k *clusters and* $\widehat{\mu}_j^\star$ *is the cluster-specific m.l.e.*

Proof. See appendix.

3.2. Multigroup comparisons.

The previous discussion can be easily extended to a generalization of the ODP statistics for the general k samples comparison problem (Storey, 2007a). We assume that data for each experimental unit i, $i = 1, \ldots, m$, are arranged by k distinct groups, and we wish to test if 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_i^l, $l = 1, \ldots, k$, $i = 1, \ldots, m$. Alternatively, data in each group may be reduced to statistics $z_i^l \sim f(z_i^l; \mu_i^l)$, and we can write $z = \{z_1, \ldots, z_m\}$, $z_i = \{z_i^1, \ldots, z_i^k\}$, 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 hypothesis for the multigroup comparison can be formalized as $H_0 : (\mu_i^1, \mu_i^2) \in A$ 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

$$\begin{aligned} z_i^l|\mu_i^l &\overset{ind}{\sim} f(z_i^l; \mu_i^l), \quad l = 1, 2,\ i = 1, \ldots, m \\ \mu_i = \{\mu_i^1, \mu_i^2\}|G &\overset{i.i.d}{\sim} G, \quad i = 1, \ldots, m \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

we can proceed as in section 2.3 and approximate the posterior probability for the i-th comparison by

$$v_i \approx \frac{\int_{A^c} f(z;\ \mu)\, dF_m(\mu)}{\int f(z;\ \mu)\, dF_m(\mu)} = \frac{\sum_{i:(\widehat{\mu}_i^1, \widehat{\mu}_i^2) \in A^c} f(z_i^1|\widehat{\mu}_i^1)\, f(z_i^2|\widehat{\mu}_i^2)}{\sum_{i=1}^m f(z;\ \widehat{\mu}_i)}, \tag{14}$$

where $\widehat{\mu}_i^1$, $\widehat{\mu}_i^2$ and $\widehat{\mu}_i$ are appropriate estimates of the relevant parameters within and across conditions. Expression (14) is an estimated ODP statistics for the multicomparison problem, as discussed in Storey et al. (2007b). As before, substituting cluster-specific estimates $\widehat{\mu}^{\star}_{s_i}$ for a selected partition s defines the corresponding BDP rule.

4. Comparison of $S_{\rm ODP}$ versus $S_{\rm BDP}$.

4.1. A Simulation Study.

We conduct a simulation study to compare $S_{\rm ODP}$ with the NP Bayesian approximation. The setup of the simulation mimicks the simulation study reported in Storey et al. (2007b, Figure 2). We assume $m = 48$ 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 that half of the observations are drawn from the null, while the other half are sampled in equal proportions from alternatives with means $\mu_i \in \{-1, 1, 2, 3\}$. We use $A = \{0\}$, and $h_{A^c}(\cdot) = N(0, 1)$. We simulated 1000 datasets. For each simulated data set we ran 2000 iterations of a posterior MCMC sampler (with 1000 iterations burn in). The results confirm the obervation in Storey et al. (2007b) that the $S_{\rm ODP}$ outperforms the UMP unbiased procedure in all cases where the alternative means are not arranged in a perfectly symmetric fashion around zero. The $S_{\rm BDP}$ further improves on the $S_{\rm ODP}$ by borrowing strength across comparisons with the multiple shrinkage induced by the DP clustering. Figure (1) shows that for any threshold of expected FP, the expected TP is comparable and slightly better under the $S_{\rm BDP}$ than under the $S_{\rm ODP}$. 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 to the $S_{\rm BDP}$ are computed with the highest posterior configuration and a random (last) configuration. The differences in Figure 1 are small. However, for many applications with massive multiple comparisons the number of comparisons might be a factor 10000 and more larger, leading to correspondingly larger differences in absolute numbers of true positives. Most importantly, the improvements come at no additional experimental cost. See the example in Section 4.2.

We also considered a similar comparison using (11) with posterior means substituted for $\widehat{\mu}_i$, and using $A = (-\varepsilon, \varepsilon)$, for several (small) values of ε, instead of the point null. The plot (not shown) of expected TP versus FP showed no substantial differences to Figure 1.

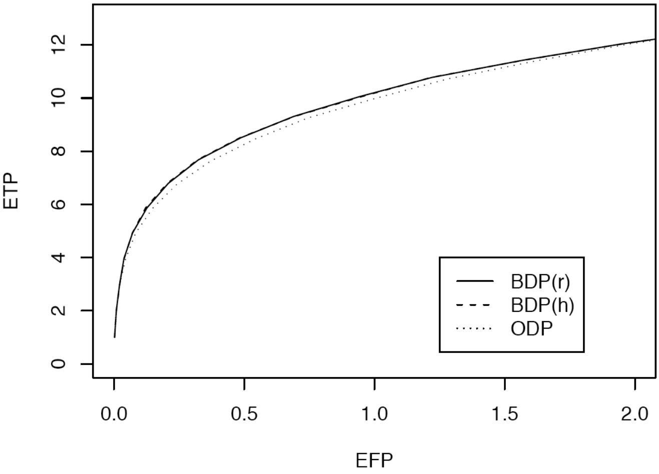


Fig. 1. A comparison in terms of ETP versus of EFP of the S_{ODP} and the S_{BDP}. $BDP(h)$ refers to the S_{BDP} computed for the cluster configuration corresponding to the maximum posterior value observed in the MCMC iterations; $BDP(r)$ refers to a random (last) configuration. See 4.1 for details.

4.2. A Microarray Data Example

We compare $\hat{S}_{\text{ODP}}$ and S_{BDP} by analyzing microarray data from breast cancer tumor tissues. The data have been analyzed, among others, in Hedenfalk et al (2001), Storey and Tibshirani (2003) and Storey et al. (2007b) and can be obtained from http://research.nhgri.nih.gov/microarray/NEJM_Supplement/. The data consist of 3,226 gene expression measurements on $n_1 = 7$ BRCA1 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. Eventually, we were left with $m = 3,169$ genes.

Let x_{ij} be the log_2 expression measurement for gene i on array j, $i = 1, \ldots, m$, $j = 1, \ldots, n$. We test differential expression between BRCA1 and BRCA2 mutation genes using the two samples t statistics $t_i = (\overline{x}_{i1} - \overline{x}_{i2})/\sqrt{(s_{i1}^2/n_1 + s_{i2}^2/n_2)}$, where $\overline{x}_{i,k}$ 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 are interested in testing $H_0 : t_i \sim N(0, \sigma)$. For simplicity, we fix $\sigma = 1$.

Our purpose is to assess the relative performance of the estimated $\hat{S}_{\text{ODP}}$ versus the NP Bayes rule approximation (11). For a fair comparison, we evaluate the S_{BDP} as a frequentist rule with rejection region $\{S_{\text{BDP}} \geq c\}$. The power of the test is evaluated in terms of the FDR and the q-value. See Storey (2002) and Storey et al. (2007b) for more discussion of the q-value and its evaluation. We briefly summarize the details that are relevant for the present comparison.

The q-value is defined for a sequence of tests that are characterized by a nested sequence of significance regions. The q-value for the i-th comparison is the minimum expected proportion of false positives incurred when rejecting H_{0i}. It is the natural pFDR analogue of the p-value. Evaluation of the q-value typically requires simulation under the null-hypothesis. A model for residuals $\varepsilon_{ij} = x_{ij} - \hat{\mu}_{i,k(j)}$ is obtained, where $k(j)$ denotes which group array j belongs to, and $\hat{\mu}_{ik} = \sum_{\ell:\, k(\ell)=j} x_{i\ell}/n_k$, where $n_k = \sum I(k(i) = j)$. The estimated residuals are then added to the overall gene-specific mean $\hat{\mu}_{i0} = \sum_{j=1}^{n} x_{ij}/n$, so that the result can be regarded as a sample from the null. A Monte Carlo sample of estimated null statistics $\hat{S}_{\text{ODP}}$ (and S_{BDP}) are then evaluated by a bootstrap resampling scheme from the transformed arrays. The Monte Carlo sample is used to compute the EFP and ETP for any given threshold c. The evaluation of S_{BDP} is based on 2000 iterations of a posterior MCMC sampler (1000 burn in). For the bootstrap step, we obtained 1000 samples. The number of significant genes detected at each q-value is shown for the $\hat{S}_{\text{ODP}}$ and the S_{BDP} in Figure 2. We report the S_{BDP} as computed on the basis of the cluster configuration of a single iteration of the MCMC sampling scheme, assumed in its stationary condition. Specifically, we use the partition from a random iteration and from the iteration that yields the maximum posterior probability. Other choices are possible. However, our experience does not suggest significantly different conclusions using alternative choices. In Figure 2 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 (4.2), 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 revealed by the ODP and 47 by the standard FDR procedure devised by Benjamini and Hochberg (1995). Out of the 11 additional genes, 7 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.

5. Extensions of the ODP and BDP

5.1. Weighted Loss Functions.

Once the optimality criterion and the probability model that lead to the ODP are identified, it is easy to modify the procedure to better accomodate preferences and losses other than (5). Often some discoveries might be considered more important than others. For example, if the experimenter's interest is in testing deviations from a common mean (e.g., $A = \{\mu_i = 0\}$), one might be more interested in large deviations from the common mean. 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) describe a decision theoretic approach to multicomparison experiments, based on the specification

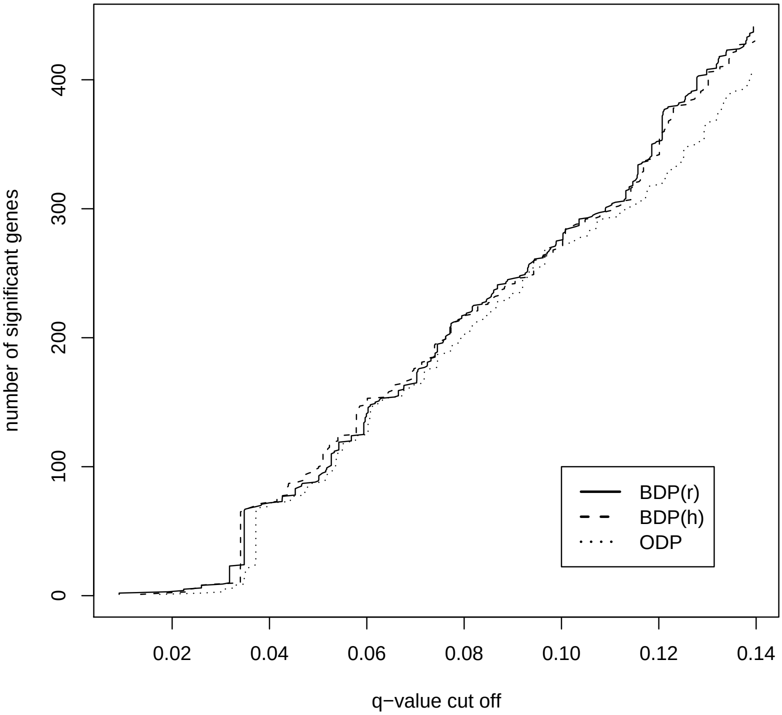


Fig. 2. A comparison of the S_{BDP} and the S_{ODP} for identifying differentially expressed data (see section 4.2). The number of genes found to be significant over a range of estimated q-value cut-offs is showns. The $BDP(h)$ curve corresponds to the S_{BDP} computed by using the cluster configuration leading to the highest posterior in the MCMC sampler, while the $BDP(r)$ refers to the S_{BDP} computed by using a random configuration in the MCMC; any choice being equivalent, here we show the result for the last configuration.

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%

Fig. 3. The intersection between the decisions with the ODP and the BDP procedure.

of separate loss functions for false positives and false negatives. Similarly, we consider the following loss function,

$$L(d, \theta, z) = -\sum d_i \; t(\mu_i) + \lambda \sum d_i \, (1 - r_i), \tag{15}$$

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

$$d_i^{m\star} = I(\lambda v_i + \overline{t(\mu_i)} > t), \tag{16}$$

for some threshold t. Although the nature of the rule has changed, we can still proceed as before and define a modified S_{ODP} statistics that approximates the Bayes rule. Let $\overline{t(\mu_i)} \approx \sum_{j=1}^m t(\widehat{\mu}_j) \, f(z_i; \, \widehat{\mu}_j) \Big/ \sum_{j=1}^m f(z_i; \, \widehat{\mu}_j)$ be an empirical Bayes estimate of $\overline{t(\mu_i)}$, justified similarly to the approximation in (8).

As before the point estimates $\widehat{\mu}_i$ could be cluster-specific m.l.e.'s $\widehat{\mu}^\star_{s_i}$, for some partition s. By an argument similar to before we can justify the following single thresholding statistic as an approximation of (16):

$$S^m_{\text{BDP}}(z_i) = \frac{\lambda \sum_{\widehat{\mu}_j \notin A} f(z_i; \, \widehat{\mu}_j) + \sum_j t(\widehat{\mu}_j)) \, f(z_i; \, \widehat{\mu}_j)}{\sum_j f(z_i; \, \widehat{\mu}_j)}. \tag{17}$$

We use $S_{\text{BDP}}^{m}(y_i)$ as a single thresholding function for the multiple comparison test in lieu of the 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) describe a setup where RT-PCR (reverse transcription polymerase chain reaction) experiments are carried out to confirm discoveries proposed by an initial microarray group comparison. Abruzzo et al. (2005) report specific values for correlations across the platforms, error variances etc. On the basis of this setup Müller et al. (2007) consider a loss function that formalizes the consequences of this follow-up experiment. The loss function includes a sampling cost for the RT-PCR experiment and a reward that is realized if the RT-PCR 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 $(\overline{\mu}_i, s_i)$ denote the posterior mean and standard deviation of the difference in mean expression for gene i between the two experimental groups. Let $\overline{\rho}, \rho^{\star}, p_{\rho}$ denote known parameters of the joint distribution of the microarray gene expression and the outcome of the RT-PCR 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 $\mu_i^A = \overline{\rho}\,(\overline{\mu}_i - s_i)$, we find a minimum sample size for the follow up experiment

$$n_i(z_i) = 2\left[(q_\alpha + q_\beta)/\mu_i^A\right]^2.$$

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

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

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

$$L(d,\theta,z) = \sum_{d_i=1} \left[-c_1 \pi_i(y_i) + n_i(y_i)\right] + c_2 \sum d_i.$$

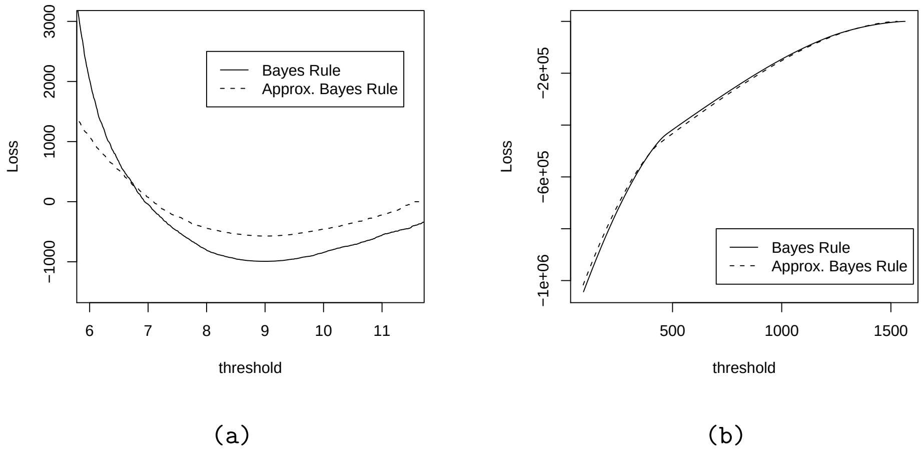
Here c_2 is a fixed cost per gene for setting up a follow-up experiment, c_1 is the (large) reward for a significant outcome in the follow-up experiment, and $c_3 \equiv 1$ is the sampling cost per gene and experiment. The Bayes rule is $d_i^{p\star} = I(c_1\pi_i - n_i \geq c_2)$. As before we can use $\overline{\mu}_i \approx \sum_{j=1}^m \widehat{\mu}_j\, f(z_i;\ \widehat{\mu}_j) \Big/ \sum_{j=1}^m f(z_i;\ \widehat{\mu}_j)$ and $s_i^2 \approx \sum_{j=1}^m (\overline{\mu}_i - \widehat{\mu}_j)^2 f(y_i \mid \widehat{\mu}_j) \Big/ \sum_{j=1}^m f(y_i;\ \widehat{\mu}_j)$ and approximate the Bayes rule by a modified ODP style statistic. Let $\widehat{\pi}_i$ and $\widehat{n}_i$ denote π_i and n_i evaluated with the approximations for $\overline{\mu}_i$ and s_i^2. We consider the modified ODP threshold statistic

$$S^p_{\text{BDP}}(y_i) = c_1 \widehat{\pi}_i(y_i) - \widehat{n}_i$$

Figure 4 compares the exact Bayes rules (16) and $d_i^{p\star} = I(c_1\pi_i - n_i \geq c_2)$ with the tests based on the approximate ODP statistic S^m_{BDP} and S^p_{BDP}, respectively.

5.2. 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 (16) remain meaningful under any probability model, as long as v_i and $\overline{t(\mu_i)}$ have meaningful interpretations. In particular, the prior probability model must give positive probability to the null hypothesis, and the posterior mean of $t(\mu_i)$ must exist. Subject to this



(a)

(b)

Fig. 4. Expected loss under the exact Bayes rules $d^{m\star}$ (left panel) and $d^{p\star}$ (right panel) plotted against the cutoff t (solid lines). The dashed lines show the expected loss under the ODP rules based on S^m_{BDP} (left panel) and S^p_{BDP} (right panel).

minimal constraint, any probability model can be used, including probability models with complicated dependence structure such as spatial models.

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., 2007). Similarly, in the analysis of fMRI 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) propose 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.

We consider a variation of the ODP suitable for spatial inference problems, using a specific spatial probability model as an example. We use the spatial model proposed by Gelfand et al. (2005). Let $\{Y(s), s \in D\}$ be a random field, where $D \subset R^d$, $d \geq 2$. Let $s^{(m)} = (s_1, ..., s_m)$ 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 $= \{Y_i(s_1), ..., Y_i(s_n)\}^T$, i = 1,...,m. We assume

$$Y_i \mid \theta_i \overset{ind}{\sim} f(y_i \mid \mu + \theta_i), \quad i = 1, \ldots, m \tag{18}$$

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

$$\begin{aligned} \theta_i \mid G &\overset{iid}{\sim} G \quad i = 1, \ldots, m \\ G &\sim DP(\alpha, G_0), \end{aligned} \tag{19}$$

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 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 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 the loss (5) and the 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 (18)-(19), it is possible to use the asymptotic arguments detailed in section 2.3 and define a BDP statistics for the spatial testing problem.

The 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 into the loss function a penalty for such irregularities, i.e.

$$L(d, \boldsymbol{\theta}, \mathbf{y}) = -\sum d_j\, r_j + \lambda \sum d_j\,(1 - r_j) + \gamma\, PI, \tag{20}$$

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 (20) is

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

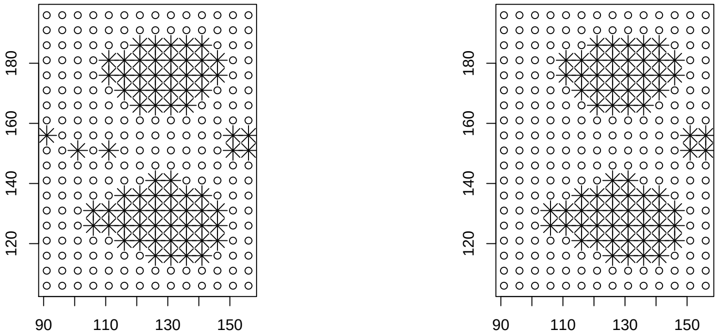
Finding $d^\star$ requires numerical optimization.

We illustrate the approach with a dataset of 18 individuals who underwent an MRI scan to detect signals of neurodegenerative patterns typical of the Alzheimer's disease (Ashburner et al., 2003). The data have been provided by the Laboratory of Epidemiology and Neuroimaging, IRCSS, Centro San Giovanni di Dio, Brescia, Italy and have been previously normalized with the freely available SPM5 sowftare (http://www.fil.ion.ucl.ac.uk/spm/, see Friston et al. (1995) and Worsley and Friston (1995)). For simplicity, the dataset is restricted to gray density matter intensity values collected on a regular two-dimensional grid of 14 × 19 pixels encompassing the hippocampus and are treated as continuous. The data have been analyzed in Petrone et al. (2008) before, although with a different inferential aim.

We assume the random effect model (18)-(19), where f is a multivariate gaussian, with mean $\mu + \theta$ and covariance matrix $\tau^2 I_n$. The base measure G_0 is a mean-zero 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, (18)-(19) defines a DP mixture of spatial processes (Gelfand et al., 2005). The model is sufficiently flexible to account for most of the spatial dependence observed in each individual. However, it is known that one of the marks of the 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 introduce several kinds of penalties into (20) 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 points for which $d_i(s) = 1$.

We use a numerical procedures to explore the action space and minimize (20). We find the optimal decision $d^\star$ by a random search, initialized with the optimal rule under $\gamma = 0$.

Figure 5.2 shows the resulting optimal rule for one individual in the MRI dataset. We are interested in detecting regions of low gray 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 dataset. The activation threshold for the posterior probability is $t = 0.8$. Figure 5.2 shows the activation map for an individual with recognizable signs of hypoccampal atrophy for $\gamma = 0$ (panel (a)) and for $\gamma = \frac{1}{2}\lambda$ (panel (b)). The additional penalty term provided a principled and coherent means of removing the singleton clusters that would otherwise be reported.



(a) Optimal decision $d^\star$ under $\gamma = 0$ (b) Optimal decision $d^\star$ with $\gamma = \frac{1}{2}\lambda$

Fig. 5. The effect of a loss function disfavoring isolated signals on the decisions taken according to loss (5). See 5.2 for details.

6. Conclusions and Summary

Starting from an interpretation of the ODP as an approximate Bayes rule we 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. The improvement is small, but comes at little additional cost. 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 as the original loss function the resulting rule can still be approximated by a single thresholding statistic similar to the ODP.

The use of a decision theoretic framework provides a convenient assurance of coherent inference for the proposed approach. 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 inference goals. Often that is sufficient to obtain a reasonable rule. But we still caution to critically validate and if necessary revise the inference in the light of evaluations such as frequentist operating characteristics. Also, the proposed methods are more computation 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 m.l.e.'s.

The main strengths of the proposed approach are the generality and the assurance of coherent inference. The approach is general in the sense that the proposed methods 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 proposed approach is the need and the opportunity to explicitely think about the inference goals and 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.

Appendix.

Proof of Preposition 1

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

$$\text{pFDR} = E\left(\frac{FP}{D} \mid D > 0\right) = E\left(\frac{\sum_{i=1}^n d_i(1-r_i)}{\sum_{i=1}^n d_i} \mid D > 0\right). \tag{21}$$

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,

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

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

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

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

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

Proof of Theorem 2

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

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

Both numerator and denominator take the form of

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

for a Borel set B. Let $s^{o(m)}$ be a vector of cluster indicators, that is $s_i^o = j$ iff $\mu_i = \mu_j^o$, $i = 1,\ldots,m$, $j = 1,\ldots,k$.

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

$$E\left(\int_B p(z_m \mid \mu_m)\,G_k(d\mu_m) \mid z_1,\ldots,z_{m-1}\right) = \frac{m}{\alpha+m}\sum_{s^{(m-1)}} p(s^{(m-1)} \mid z_1,\ldots,z_{m-1}) \times$$
$$\times \sum_{j=1}^k \frac{\frac{\alpha}{K}+m_j^o}{m}\int_B p(z_m \mid \mu_j^o)\,p(d\mu_j^o \mid z_i : s_i = j, i = 1,\ldots,m-1), \tag{22}$$

where $m_j^o = card\{s_i^o : s_i^o = j, i = 1,\ldots,m-1\}$ is the number of elements in cluster j. If $m_j^o = 0$,then $p(d\mu_j^o \mid z_i : s_i = j) \equiv G^*(d\mu_j^o)$. Expression (22) highlights that any partition of $\{1,\ldots,m\}$ can be obtained from a corresponding partition of $\{1,\ldots,m-1\}$ by adding the m-th observation to any of the previous cluster or forming a new one.

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

$$\int_B p(z_m \mid \mu_j^o)\, p(d\mu_j^o \mid z_i : s_i^o = j) \approx p(z_m \mid \widehat{\mu}_j^o)\, \Phi(\widehat{\mu}_j^o \in B)[1 + O(m_j^{o\,-2})],$$

where $\widehat{\mu}_j^o$ is the MLE estimate 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 $\{\mu_j^o;\ m_j^o > 0\}$ as the set of unique values $\{\mu_j^\star,\ j = 1, \ldots, L\}$.

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

References

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. *The Lancet Neurology*, **2**, 79–88.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discover rate – a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **75**, 289–300.

Cohen, A. and Sackrowitz, H. (2007) More on the inadmissibility of step-up. *Journal of Multivariate Analysis*, **98**, 481–492.

Dahl, D. and Newton, M. (2007) Multiple hypothesis testing by clustering treatment effects. *J. Am. Statistic. Assoc.*, **102**, 517–526.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Statist. Assoc.*, **96**, 1151–1160.

Ferguson, T. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.

— (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), 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. *Humain Brain Mapping*, **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.* Springer.

Gopalan, R. and Berry, D. (1993) Bayesian multiple comparisons using dirichlet process priors. *J. Am. Statist. Assoc.*, **93**, 1130–1139.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., M., E., 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, Olli-P., Borg, A. and Trent, J. (2001) Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, **344**, 539–549.

Ishwaran, H. and James, L. (2003) Some further developments for stick-breaking priors: finite and infinite clustering and classification. *Sankhya Series A*, **65**, 577–592.

Lo, A. (1984) On a class of bayesian nonparametric estimates: I density estimates. *Ann. Statist.*, **12 (1)**, 351–357.

Müller, P., G., P. and K., R. (2007) Fdr and bayesian multiple comparisons rules. In *Bayesian Statistics 8* (eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, Heckerman, A. D., Smith and M. West). Oxford, UK: Oxford University Press.

Neal, R. M. (2000) Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Pacifico, M. P., Genovese, C., Verdinelli, I. and Wasserman, L. (2004) False discovery control for random fields. *J. Am. Statistic. Assoc.*, **99**, 1002–1014.

Petrone, S., Guindani, M. and Gelfand, A. (2008) Hybrid Dirichlet processes for functional data. *Under invited revision to J. R. Statist. Soc. B.*

Rodriguez, A., Dunson, D. and Gelfand, A. (2008) The nested Dirichlet process. *J. Am. Statist. Ass.*

Schervish, M. J. (1995) *Theory of Statistics.* Springer.

Scott, J. and Berger, J. (2003) An exploration of aspects of bayesian multiple testing. *Journal of Statistical Planning and Inferece*, **136**, 2144–2162.

Storey, J. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. B*, **64**, 479–498.

— (2007a) The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc. B*, **69 (3)**, 347–368.

Storey, J., Dai, J. and Leek, J. (2007b) 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 (16)**, 9440–9445.

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. (2007) Loss function approaches to predict a spatial quantile and its exceedance region. *Submitted.*