Chapter 7

Bayesian Nonparametric Approach to Multiple Testing

Subhashis Ghosal[1] and Anindya Roy[2]

[1]*Department of Statistics,*
North Carolina State University,
2501 Founders Drive,
Raleigh NC 27695-8203, USA
sghosal@stat.ncsu.edu

[2]*Department of Mathematics & Statistics,*
University of Maryland Baltimore County,
1000 Hilltop Circle,
Baltimore, MD 21250, USA
anindya@math.umbc.edu

Motivated by the problems in genomics, astronomy and some other emerging fields, multiple hypothesis testing has come to the forefront of statistical research in the recent years. In the context of multiple testing, new error measures such as the false discovery rate (FDR) occupy important roles comparable to the role of type I error in classical hypothesis testing. Assuming that a random mechanism decides the truth of a hypothesis, substantial gain in power is possible by estimating error measures from the data. Nonparametric Bayesian approaches are proven to be particularly suitable for estimation of error measure in multiple testing situation. A Bayesian approach based on a nonparametric mixture model for p-values can utilize special features of the distribution of p-values that significantly improves the quality of estimation. In this paper we describe the nonparametric Bayesian modeling exercise of the distribution of the p-values. We begin with a brief review of Bayesian nonparametric concepts of Dirichlet process and Dirichlet mixtures and classical multiple hypothesis testing. We then review recently proposed nonparametric Bayesian methods for estimating errors based on a Dirichlet mixture of prior for the p-value density. When the test statistics are independent, a mixture of beta kernels can adequately model the p-value density, whereas in the dependent case one can consider a Dirichlet mixture of multivariate skew-normal kernel prior for probit transforms of

the p-values. We conclude the paper by illustrating the scope of these methods in some real-life applications.

7.1. Bayesian Nonparametric Inference

To make inference given an observed set of data, one needs to model how the data are generated. The limited knowledge about the mechanism often does not permit explicit description of the distribution given by a relatively few parameters. Instead, only very general assumptions leaving a large portion of the mechanism unspecified can be reasonably made. This nonparametric approach thus avoids possible gross misspecification of the model, and understandably is becoming the preferred approach to inference, especially when many samples can be observed. Nonparametric models are actually not parameter free, but they contain infinite dimensional parameters, which can be best interpreted as functions. In common applications, the cumulative distribution function (c.d.f.), density function, nonparametric regression function, spectral density of a time series, unknown link function in a generalized linear model, transition density of a Markov chain and so on can be the unknown function of interest. Classical approach to nonparametric inference has flourished throughout the last century. Estimation of c.d.f. is commonly done by the empirical c.d.f., which has attractive asymptotic properties. Estimation of density, regression function and similar objects in general needs smoothing through the use of a kernel or through a basis expansion. Testing problems are generally approached through ranks, which typically form the maximal invariant class under the action of increasing transformations.

Bayesian approach to inference offers a conceptually straightforward and operationally convenient method, since one needs only to compute the posterior distribution given the observations, on which the inference is based. In particular, standard errors and confidence sets are automatically obtained along with a point estimate. In addition, the Bayesian approach enjoys philosophical justification and often Bayesian estimation methods have attractive frequentist properties, especially in large samples. However, Bayesian approach to nonparametric inference is challenged by the issue of construction of prior distribution on function spaces. Philosophically, specifying a genuine prior distribution on an infinite dimensional space amounts to adding infinite amount of prior information about all fine details of the function of interest. This is somewhat contradictory to the motivation of nonparametric modeling where one likes to avoid specifying too much

about the unknown functions. This issue can be resolved by considering the so called "automatic" or "default" prior distributions, where some tractable automatic mechanism constructs most part of the prior by spreading the mass all over the parameter space, while only a handful of key parameters may be chosen subjectively. Together with additional conditions, large support of the prior helps the posterior distribution concentrate around the true value of the unknown function of interest. This property, known as posterior consistency, validates a Bayesian procedure from the frequentist view, in that it ensures that, with sufficiently large amount of data, the truth can be discovered accurately and the data eventually overrides any prior information. Therefore, a frequentist will be more likely to agree to the inference based on a default nonparametric prior. Lack of consistency is thus clearly undesirable since this means that the posterior distribution is not directed toward the truth. For a consistent posterior, the speed of convergence to the true value, called the rate of convergence, gives a more refined picture of the accuracy of a Bayesian procedure in estimating the unknown function of interest.

For estimating an arbitrary probability measure (equivalently, a c.d.f.) on the real line, with independent and identically distributed (i.i.d.) observations from it, Ferguson ([19]) introduced the idea of a Dirichlet process — a random probability distribution P such that for any finite measurable partition $\{B_1, \ldots, B_k\}$ of $\mathbb{R}$, the joint distribution of $(P(B_1), \ldots, P(B_k))$ is a finite dimensional Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_k))$, where α is a finite measure called the base measure of the Dirichlet process $\mathcal{D}_\alpha$. Since clearly $P(A) \sim \text{Beta}(\alpha(A), \alpha(A^c))$, we have $\text{E}(P(A)) = \alpha(A)/(\alpha(A) + \alpha(A^c)) = G(A)$, where $G(A) = \alpha(A)/M$, a probability measure called the center measure and $M = \alpha(\mathbb{R})$, called the precision parameter. This implies that if $X|P \sim P$ and $P \sim \mathcal{D}_\alpha$, then marginally $X \sim G$. Observe that $\text{var}(P(A)) = G(A)G(A^c)/(M+1)$, so that the prior is more tightly concentrated around its mean when M is larger. If P is given the measure $\mathcal{D}_\alpha$, we shall write $P \sim \text{DP}(M, G)$. The following give the summary of the most important facts about the Dirichlet process:

(i) If $\int |\psi| dG < \infty$, then $\text{E}(\int \psi dP) = \int \psi dG$.
(ii) If $X_1, \ldots, X_n | P \overset{iid}{\sim} P$ and $P \sim \mathcal{D}_\alpha$, then $P|X_1, \ldots, X_n \sim \mathcal{D}_{\alpha + \sum_{i=1}^n \delta_{X_i}}$.
(iii) $\text{E}(P|X_1, \ldots, X_n) = \frac{M}{M+n} G + \frac{n}{M+n} \mathbb{P}_n$, a convex combination of the prior mean and the empirical distribution $\mathbb{P}_n$.
(iv) Dirichlet sample paths are a.s. discrete distributions.
(v) The topological support of $\mathcal{D}_\alpha$ is $\{P^* : \text{supp}(P^*) \subset \text{supp}(G)\}$.

(vi) The marginal joint distribution of $(X_1, \ldots, X_n)$ from P, where $P \sim \mathcal{D}_\alpha$, can be described through the conditional laws

$$X_i|(X_l, l \neq i) \sim \begin{cases} \delta_{\phi_j}, & \text{with probability } \frac{n_j}{M+n-1},\ j = 1, \ldots, k_{-i}, \\ G, & \text{with probability } \frac{M}{M+n-1}, \end{cases}$$

where k_{-i} is the number of distinct observations in X_l, $l \neq i$ and $\phi_1, \ldots, \phi_{k_{-i}}$ are those distinct values with multiplicities $n_1, \ldots, n_{k_{-i}}$. Thus the number of distinct observations K_n in $X_1, \ldots, X_n$, is generally much smaller than n with $\mathrm{E}(K_n) = M \sum_{i=1}^{n} (M + i - 1)^{-1} \sim M \log(n/M)$, introducing sparsity.

(vii) Sethuraman's ([51]) stick-breaking representation: $P = \sum_{i=1}^{\infty} V_i \delta_{\theta_i}$, where $\theta_i \overset{iid}{\sim} G$, $V_i = [\prod_{j=1}^{i-1}(1 - Y_j)]Y_i$, $Y_i \overset{iid}{\sim} \mathrm{Beta}(1, M)$. This allows us to approximately generate a Dirichlet process and is indispensable in various complicated applications involving the Dirichlet process, where posterior quantities can be simulated approximately with the help of a truncation and Markov chain Monte-Carlo (MCMC) techniques.

In view of (iii), clearly G should be elicited as the prior guess about P, while M should be regarded as the strength of this belief. Actual specification of these are quite difficult in practice, so we usually let G contain additional hyperparameters ξ, and some flat prior is put on ξ, leading to a mixture of Dirichlet process ([1]).

A widely different scenario occurs when one mixes parametric families nonparametrically. Assume that given a latent variable θ_i, the observations X_i follows a parametric density $\psi(\cdot; \theta_i)$, $i = 1, \ldots, n$, respectively, and the random effects $\theta_i \overset{iid}{\sim} P$, $P \sim \mathcal{D}_\alpha$ ([20], [33]). In this case, the density of the observation can be written as $f_P(x) = \int \psi(x; \theta) dP(\theta)$. The induced prior distribution on f_P through $P \sim \mathrm{DP}(M, G)$ is called a Dirichlet process mixture (DPM). Since $f_P(x)$ is a linear functional of P, the expressions of posterior mean and variance of the density $f_P(x)$ can be analytically expressed. However, these expressions contain enormously large number of terms. On the other hand, computable expressions can be obtained by MCMC methods by simulating the latent variables $(\theta_1, \ldots, \theta_n)$ from their posterior distribution by a scheme very similar to (vi); see [18]. More precisely, given θ_j, $j \neq i$, only X_i affects the posterior distribution of θ_i. The observation X_i weighs the selection probability of an old θ_j by $\psi(X_i; \theta_j)$, and the fresh draw by $M \int \psi(X_i; \theta) dG(\theta)$, and a fresh draw, whenever obtained, is taken from the "baseline posterior" defined by

$dG_i(\theta) \propto \psi(X_i;\theta)dG(\theta)$. The procedure is known as the generalized Polya urn scheme.

The kernel used in forming DPM can be chosen in different ways depending on the sample space under consideration. A location-scale kernel is appropriate for densities on the line with unrestricted shape. In Section 7.3, we shall use a special type of beta kernels for decreasing densities on the unit interval modeling the density of p-values in multiple hypothesis testing problem.

To address the issue of consistency, let Π be a prior on the densities and let f_0 stand for the true density. Then the posterior probability of a set B of densities given observations $X_1, \ldots, X_n$ can be expressed as

$$\Pi(f \in B|X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}{\int \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}. \tag{7.1}$$

When B is the complement of a neighborhood U of f_0, consistency requires showing that the expression above goes to 0 as $n \to \infty$ a.s. $[P_{f_0}]$. This will be addressed by showing that the numerator in (7.1) converges to zero exponentially fast, while the denominator multiplied by $e^{\beta n}$ goes to infinity for all $\beta > 0$. The latter happens if $\Pi(f : \int f_0 \log(f_0/f) < \epsilon) > 0$ for all $\epsilon > 0$. The assertion about the numerator in (7.1) holds if a uniformly exponentially consistent test exists for testing the null hypothesis $f = f_0$ against the alternative $f \in U^c$. In particular, the condition holds automatically if U is a weak neighborhood, which is the only neighborhood we need to consider in our applications to multiple testing.

7.2. Multiple Hypothesis Testing

Multiple testing procedures are primarily concerned with controlling the number of incorrect significant results obtained while simultaneously testing a large number of hypothesis. In order to control such errors an appropriate error rate must be defined. Traditionally, the family-wise error rate (FWER) has been the error rate of choice until recently when the need was felt to define error rates that more accurately reflect the scientific goals of modern statistical applications in genomics, proteomics, functional magnetic resonance imaging (fMRI) and other biomedical problems. In order to define the FWER and other error rates we must first describe the different components of a typical multiple testing problem. Suppose $\mathrm{H}_{10}, \ldots, \mathrm{H}_{m0}$ are m null hypotheses whose validity is being tested simultaneously. Suppose m_0 of those hypotheses are true and after making

Table 7.1. Number of hypotheses accepted and rejected and their true status.

Decision			
Hypothesis	Accept	Reject	Total
True	U	V	m_0
False	T	S	$m - m_0$
Total	Q	R	m

decisions on each hypothesis, R of the m hypotheses are rejected. Also, denote the m ordered p-values obtained from testing the m hypotheses as $X_{(1)} < X_{(2)} < \cdots < X_{(m)}$. Table 7.1 describes the components associated with this scenario.

The FWER is defined as the probability of making at least one false discovery, i.e. FWER $= P(V \geq 1)$. The most common FWER controlling procedure is the Bonferroni procedure where each hypotheses is tested at level α/m to meet an overall error rate of α; see [35]. When m is large, this measure is very conservative and may not yield any "statistical discovery", a term coined by [54] to describe a rejected hypothesis. Subsequently, several generalization of the Bonferroni procedure were suggested where the procedures depend on individual p-values, such as [52], [30], [31], [29] and [42]. In the context of global testing where one is interested in the significance of a set of hypotheses as a whole, [52] introduced a particular sequence of critical values, $\alpha_i = i\alpha/n$, to compare with each p-value. More recently, researchers proposed generalization of the FWER (such as the k-FWER) that is more suitable for modern applications; see [32].

While the FWER gives a very conservative error rate, at the other extreme of the spectrum of error rates is the per comparison error rate (PCER) where significance of any hypothesis is decided without any regard to the significance of the rest of the hypothesis. This is equivalent to testing each hypothesis at a fixed level α and looking at the average error over the m tests conducted, i.e. PCER $= \mathrm{E}(V/m)$. While the PCER is advocated by some ([53]) it is too liberal and may result in several false discoveries. A compromise was proposed by [7] where they described a sequential procedure to control the false discovery rate (FDR), defined as FDR $= \mathrm{E}(V/R)$. The ratio V/R is defined to be zero if there are no rejections. The FDR as an error rate has many desirable properties. First of all, as described in [7] and by many others, one can devise algorithms to control FDR in multiple testing situation under fairly general joint behavior of the test statistics for the hypotheses. Secondly, if all hypotheses are true, controlling FDR

is equivalent to controlling the FWER. In general, FDR falls between the other two error rates, the FWER and PCER (cf. [24]).

The Benjamini-Hochberg (B-H) FDR control procedure is a sequential step-up procedure where the p-values (starting with the largest p-value) are sequentially compared with a sequence of critical values to find a critical p-value such that all hypotheses with p-values smaller than the critical value are rejected. Suppose $\hat{k} = \max\{i : X_{(i)} \leq \alpha_i\}$ where $\alpha_i = i\alpha/m$. The the B-H procedure rejects all hypotheses with p-values less than or equal to $X_{(\hat{k})}$. If no such $\hat{k}$ exists, then none of the hypotheses is rejected. Even though the algorithm sequentially steps down through the sequence of p-values, it is called a step-up procedure because this is equivalent to stepping up with respect to the associated sequence of test statistics to find a minimal significant test value. The procedure is also called a linear step-up procedure due to the linearity of the critical function α_i with respect to i. [9], [46], [59] among others have shown the FDR associated with this particular step-up procedure is exactly equal to $m_0\alpha/m$ in the case when the test statistics are independent and is less than $m_0\alpha/m$ if the test statistics have positive dependence: for every test function ϕ, the conditional expectation $\mathrm{E}[\phi(X_1, \ldots, X_m)|X_i]$ is increasing with X_i for each i. [46] has suggested an analogous step-down procedure where one fails to reject all hypotheses with p-values above a critical value α_i, that is, if $\hat{l} = \min\{i : X_{(i)} > \alpha_i\}$, none of the hypotheses associated with p-value $X_{(\hat{l})}$ and above is rejected. [46] used the same set of critical values $\alpha_i = i\alpha/m$ as in [7] which also controls the FDR at the desired level (see [47]). However, for the step-down procedure even in the independent case the actual FDR may be less than $m_0\alpha/m$.

Since in the independent case the FDR of the linear step-up procedure is exactly equal to $m_0\alpha/m$, if the proportion of true null hypotheses, $\pi = m_0/m$, is known then α can be adjusted to get FDR equal to any target level. Specifically, if $\alpha_i = i\alpha/(m\pi)$ then the FDR of the linear step-up procedure is exactly equal to α in the independent case. Unfortunately, in any realistic situation m_0 is not known. Thus, in situations where π is not very close to one, FDR can be significantly smaller than the desired level, and the procedure may be very conservative with poor power properties.

Another set of sequential FDR controlling procedures were introduced more recently, where π is adaptively estimated from the data and the critical values are modified as $\alpha_i = i\alpha/(m\hat{\pi})$. Heuristically, this procedure would yield an FDR close to $\pi\alpha\mathrm{E}(\hat{\pi}^{-1})$, and if $\hat{\pi}$ is an efficient estimator of π then the FDR for the adaptive procedure will be close to the target level

α. However, merely plugging-in an estimator of π in the expression for α_i may yield poor results due to the variability of the estimator of π^{-1}. [57] suggested using $\hat{\pi} = [m - R(\lambda) + 1]/[m(1-\lambda)]$, where $R(\lambda) = \sum \mathbb{1}\{X_i \leq \lambda\}$ is the number of p-values smaller than λ and $0 < \lambda < 1$ is a constant; here and below $\mathbb{1}$ will stand for the indicator function. Similar estimators had been originally suggested by [56]. Then for any λ, choose the sequence of critical points as

$$\alpha_i = \min\Big\{\lambda, \frac{i\alpha(1-\lambda)}{m - R(\lambda) + 1}\Big\}.$$

The adaptive procedure generally yields tighter FDR control and hence can enhance the power properties of the procedures significantly ([8], [12]). Of course, the performance of the procedure will be a function of the choice of λ. [58] suggested various procedures for choosing λ. [11] suggested choosing $\lambda = \alpha/(1+\alpha)$ and they looked at the power properties of the adaptive procedure. [50] investigated theoretical properties of these two stage procedures and [22] suggested analogous adaptive step-down procedures.

The procedures described above for controlling FDR can be thought of as fixed-error rate approach where the individual hypotheses are tested at different significance level to maintain a constant overall error rate. [57, 58] introduced the fixed-rejection-region approach where $\alpha_i = \alpha$ for all i (i.e. the rejection region is fixed). The FDR given the rejection region is estimated from the data and then α is chosen to set the estimated FDR at a predetermined level. [57] also argued that since one becomes concerned about false discoveries only in the situation where there are some discoveries, one should look at the expected proportion of false discoveries conditional on the fact that there has been some discoveries. Thus the positive false discovery rate (pFDR) is defined as $\text{pFDR} = \text{E}(V/R|R > 0)$. [57] showed that if we assume a mixture model for the hypotheses, i.e., if we can assume that the true null hypothesis are arising as a Bernoulli sequence with probability π, then the expression for pFDR reduces to

$$\text{pFDR}(\alpha) = \frac{\pi\alpha}{F(\alpha)} \tag{7.2}$$

where $F(\cdot)$ is the marginal c.d.f. of the p-values. Although it cannot be controlled in the situation when there are no discoveries, given its simple expression, pFDR is ideally suited for the estimation approach. Once an estimator for pFDR has been obtained, the error control procedure reduces to rejecting all p-values less than or equal to $\hat{\gamma}$ where

$$\hat{\gamma} = \max\{\gamma : \widehat{\text{pFDR}}(\gamma) \leq \alpha\}. \tag{7.3}$$

Storey (cf. [57]) showed that the B-H linear step-up procedure can be viewed as Storey's procedure where π is estimated by 1. Therefore, it is clear that using the procedure (7.3) will improve the power substantially unless π is actually very close to 1.

Storey (cf. [58]) also showed that the pFDR can be given a Bayesian interpretation as the posterior probability of a null hypothesis being true given that it has been rejected. This interpretation connects the frequentist and the Bayesian paradigms in the multiple testing situation. Given that p-values are fundamental quantities that can be interpreted in both paradigms, this connection in the context of a procedure based on p-values is illuminating. Several multiple testing procedures have resulted by substituting different estimators of pFDR in (7.3). Most of these procedures rely on the expression (7.2) and substitute the empirical c.d.f. for $F(\alpha)$ in the denominator. These procedures mainly differ in the way they estimate π. However, since $\pi\alpha$ is less than or equal to $F(\alpha)$, there is always a risk of violating the inequality if one estimates $F(\alpha)$ and π independently. [60] suggested a nonparametric Bayesian approach that simultaneously estimates π and $F(\alpha)$ within a mixture model framework that naturally constrain the estimators to maintain the relationship. This results in a more efficient estimator of pFDR.

The case when the test statistics (equivalently, p-values) are dependent is of course of great practical interest. A procedure that controls the FDR under positive regression dependence was suggested in [9] where the B-H critical values are replaced by $\alpha_i = \frac{i\alpha}{m\sum_{j=1}^{i} j^{-1}}$. The procedure is very conservative because the critical values are significantly smaller than the B-H critical values. [50] suggested an alternative set of critical values and investigated the performance under some special dependence structures. [21] and [17] suggested modeling the probit transform of the p-values as joint normal distribution to capture dependence among the p-values. A similar procedure to model the joint behavior of the p-values was suggested by [44] who used a mixture of skew-normal densities to incorporate dependence among the p-values. This mixing distribution is then estimated using nonparametric Bayesian techniques described in Section 7.1.

Other error measure such as the local FDR ([17]) were introduced to suit modern large dimensional datasets. While the FDR depends on the tail probability of the marginal p-value distribution, $F(\alpha)$, the local FDR depends on the marginal p-value density. Other forms of generalization can be found in ([48], [49]) and the references therein. Almost all error

measures are functionals of the marginal p-value distribution, while few have been analyzed under the possibility of dependence among the p-values. A model based approach that estimates the components of the marginal distribution of the p-values has the advantage that once accurate estimates of the components of the marginal distribution are obtained, then it is possible to estimate several of these error measures and make a comparative study. Bayesian methodologies in multiple testing were discussed in [13], [60], [27] and [44]. [26] used a weighted p-value scheme that incorporates prior information about the hypothesis in the FDR controlling procedure. Empirical Bayes estimation of FDR was discussed in [15].

A particularly attractive feature of the Bayesian approach in the multiple testing situation is its ability to attach a posterior probability to an individual null hypothesis being actually false. In particular, it is easy to predict the false discovery proportion (FDP), V/R. Let $I_i(\alpha) = \mathbb{1}\{X_i < \alpha\}$ denote that the ith hypothesis is rejected at a threshold level α and let H_i be the indicator that the ith alternative hypothesis is true. The FDP process evaluated at a threshold α (cf. [25]) is defined by

$$\mathrm{FDP}(\alpha) = \frac{\sum_{i=1}^m I_i(\alpha)(1-H_i)}{\sum_{i=1}^m I_i(\alpha) + \prod_{i=1}^m (1-I_i(\alpha))}.$$

Assuming that $(H_i, I_i(\alpha))$, $i = 1, \ldots, m$, are exchangeable, [44] showed that $\mathrm{FDR}(\alpha) = \pi b(\alpha)\mathrm{P}(\text{at least one rejection})$, where $b(\alpha)$ is the expected value of a function of the indicator functions. This implies that $\mathrm{pFDR}(\alpha) = \pi b(\alpha)$, which reduces to the old expression under independence. A similar expression was derived in [9] and also in [47]. In particular, [47] showed that the quantity $b(\alpha)/\alpha$ is the expectation of a jackknife estimator of $\mathrm{E}[(1+R)^{-1}]$.

Thus the simple formula for pFDR as $\pi\alpha/F(\alpha)$ does not hold if the p-values are dependent, but the FDP with better conditional properties, seems to be more relevant to a Bayesian. Estimating the pFDR will generally involve computing high dimensional integrals, and hence will be difficult to obtain in reasonable time, but predicting the FDP is considerably simpler. Since the Bayesian methods are able to generate from the joint conditional distribution of $(H_1, \ldots, H_m)$ given data, we can predict the FDP by calculating its conditional expectation given data.

The theoretical model for the null distribution of the p-values is U[0, 1]. The theoretical null model may not be appropriate for the observed p-values in many real-life applications due to composite null hypothesis, complicated test statistic or dependence among the datasets used

to test the multiple hypothesis. For a single hypothesis, the uniform null model may be approximately valid even for very complex hypothesis testing situations with composite null and complicated test statistics; see [4]. However, as argued by [16] and [6], if the multiple hypotheses tests are dependent then the m_0 null p-values collectively can behave very differently from a collection of independent uniform random variables. For example, the histogram of the probit transformed null p-values may be significantly skinnier than the standard normal, the theoretical null distribution of the probit p-values. [16] showed that a small difference between the theoretical null and an empirical null can have a significant impact on the conclusions of an error control procedure. Fortunately, large scale multiple testing situations provide one with the opportunity to empirically estimate the null distribution using a mixture model framework. Thus, validity of the theoretical null assumption can be tested from the data and if the observed values show significant departure from the assumed model, then the error control procedure may be built based on the empirical null distribution.

7.3. Bayesian Mixture Models for p-Values

As discussed in the previous section, p-values play an extremely important role in controlling the error in a multiple hypothesis testing problem. Therefore, it is a prudent strategy to base our Bayesian approach considering p-values as fundamental objects rather than as a product of some classical testing procedure. Consider the estimation approach of Storey ([57, 58]) discussed in the previous section. Here the false indicator H_i of the ith null hypothesis, is assumed to arise through a random mechanism, being distributed as independent Bernoulli variables with success probability $1-\pi$. Under this scenario, even though the original problem of multiple testing belongs to the frequentist paradigm, the probabilities that one would like to estimate are naturally interpretable in a Bayesian framework. In particular, the pFDR function can be written in the form of a posterior probability. There are other advantages of the Bayesian approach too. Storey's estimation method of π is based on the implicit assumption that the the density of p-values h under the alternative is concentrated near zero, and hence almost every p-value over the chosen threshold λ must arise from null hypotheses. Strictly speaking, this is incorrect because p-values bigger than λ can occur under alternatives as well. This bias can be addressed through elaborate modeling of the p-value density. Further, it is unnatural to assume that the value of the alternative distribution remains

fixed when the hypotheses themselves are appearing randomly. It is more natural to assume that, given that the alternative is true, the value of the parameter under study is chosen randomly according to some distribution. This additional level of hierarchy is easily absorbed in the mixture model for the density of p-values proposed below.

7.3.1. *Independent case: Beta mixture model for p-values*

In this subsection, we assume that the test statistics, and hence the p-values, arising from different hypotheses are independent. Then the p-values $X_1, \ldots, X_m$ may be viewed as i.i.d. samples from the two component mixture model: $f(x) = \pi g(x) + (1-\pi)h(x)$, where g stands for the density of p-values under the null hypothesis and h that under the alternative. The distribution of X_i under the corresponding null hypothesis $H_{0,i}$ may be assumed to be uniformly distributed on $[0, 1]$, at least approximately. This happens under a number of scenarios:

(i) the test statistic is a continuous random variable and the null hypothesis is simple;
(ii) in situations like t-test or F-test, where the null hypothesis has been reduced to a simple one by considerations of similarity or invariance;
(iii) if a conditional predictive p-value or a partial predictive p-value ([4], [41]) is used.

Thus, unless explicitly stated, hereafter we assume that g is the uniform density. It is possible that this assumption fails to hold, which will be evident from the departure of the empirical null distribution from the theoretical null. However, even when this assumption fails to hold, generally the actual g is stochastically larger than the uniform. Therefore it can be argued that the error control procedures that assume the uniform density remain valid in the conservative sense. Alternatively, this difference can be incorporated in the mixture model by allowing the components of the mixture distribution that are stochastically larger than the uniform distribution to constitute the actual null distribution.

The density of p-values under alternatives is not only concentrated near zero, but usually has more features. In most multiple testing problems, individual tests are usually simple one-sided or two-sided z-test, χ^2-test, or more generally, tests for parameters in a monotone likelihood ratio (MLR) family. When the test is one-sided and the test statistic has the MLR property, it is easy to see that the density of p-values is decreasing (Proposition 1

of [27]). For two-sided alternatives, the null distribution of the test statistic is often symmetric, and in that case, a two-sided analog of the MLR property implies that the p-value density is decreasing (Proposition 2 of [27]). The p-value density for a one-sided hypothesis generally decays to zero as x tends to 1. For a two-sided hypothesis, the minimum value of the p-value density will be a (small) positive number. For instance, for the two-sided normal location model, the minimum value is $e^{-n\theta^2/2}$, where n is the sample size on which the test is based on. In either case, the p-value density looks like a reflected "J", a shape exhibited by a beta density with parameters $a < 1$ and $b \geq 1$. In fact, if we are testing for the scale parameter of the exponential distribution, it is easy to see that the p-value density is exactly beta with $a < 1$ and $b = 1$. In general, several distributions on $[0, 1]$ can be well approximated by mixtures of beta distributions (see [14], [40]). Thus it is reasonable to approximate the p-value density under the alternative by an arbitrary mixture of beta densities with parameters $a < 1$ and $b \geq 1$, that is, $h(x) = \int \text{be}(x|a, b)dG(a, b)$, where $\text{be}(x; a, b) = x^{a-1}(1-x)^{b-1}/B(a, b)$ is the beta density with parameters a and b, and $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function. The mixing distribution can be regarded as a completely arbitrary distribution subject to the only restriction that G is concentrated in $(0, 1) \times [1, \infty)$. [60] took this approach and considered a Dirichlet process prior on the mixing distribution G. Note that, if the alternative values arise randomly from a population distribution and individual p-value densities conditional on the alternative are well approximated by mixtures of beta densities, then the beta mixture model continues to approximate the overall p-value density. Thus, the mixture model approach covers much wider models and has a distinct advantage over other methods proposed in the literature. The resulting posterior can be computed by an appropriate MCMC method, as described below. The resulting Bayesian estimator, because of shrinkage properties, offers a reduction in the mean squared error and is generally more stable than its empirical counterpart considered by Storey ([57, 58]). [60] ran extensive simulation to demonstrate the advantages of the Bayesian estimator.

The DPM model is equivalent to the following hierarchical model, where associated with each X_i there is a latent variable $\theta_i = (a_i, b_i)$,

$$X_i|\theta_i \sim \pi + (1 - \pi)\,\text{be}(x_i|\theta_i), \quad \theta_1, \ldots, \theta_m|G \overset{iid}{\sim} G \quad \text{and} \quad G \sim \text{DP}(M, G_0).$$

The random measure G can be integrated out from the prior distribution to work with only finitely many latent variables $\theta_1, \ldots, \theta_m$.

In application to beta mixtures, it is not possible to choose G_0 to be conjugate with the beta likelihood. Therefore it is not possible to obtain closed-form expressions for the weights and the baseline posterior distribution in the generalized Polya urn scheme for sampling from the posterior distribution of $(\theta_1, \ldots, \theta_n)$. To overcome this difficulty, the no-gaps algorithm ([34]) may be used, which can bypass the problems of evaluating the weights and sampling from the baseline posterior. For other alternative MCMC schemes, consult [36].

[60] gave detailed description of how the no-gaps algorithm can be implemented to generate samples from the posterior of $(\theta_1, \ldots, \theta_m, \pi)$. Once MCMC sample values of $(\theta_1, \ldots, \theta_m, \pi)$ are obtained, the posterior mean is approximately given by the mean of the sample π-values. Since the pFDR functional is not linear in (G, π), evaluation of the posterior mean of pFDR(α) requires generating posterior samples of the infinite dimensional parameter h using Sethuraman's representation of G. This is not only cumbersome, but also requires truncating the infinite series to finitely many terms and controlling the error resulting from the truncation. We avoid this path by observing that, when m is large (which is typical in multiple testing applications), the "posterior distribution" of G given $\theta_1, \ldots, \theta_m$ is essentially concentrated at the "posterior mean" of G given $\theta_1, \ldots, \theta_m$, which is given by $\mathrm{E}(G|\theta_1, \ldots, \theta_m) = (M+m)^{-1}MG_0 + (M+m)^{-1}\sum_{i=1}^m \delta_{\theta_i}$, where $\delta_\theta(x) = 1\!\!1\{\theta \le x\}$ now stands for the c.d.f. of the distribution degenerate at θ. Thus the approximate posterior mean of pFDR(α) can be obtained by the averaging the values of $\pi\alpha/[(M+m)^{-1}MG_0(\alpha) + (M+m)^{-1}\sum_{i=1}^m \delta_{\theta_i}(\alpha)]$ realized in the MCMC samples. In the simulations of [60], it turned out that the sensitivity of the posterior to prior parameters is minimal.

In spite of the success of the no gaps algorithm in computing the Bayes estimators of π and pFDR(α), the computing time is exorbitantly high in large scale applications. In many applications, real-time computing giving instantaneous results is essential. Newton's algorithm ([38], [39], [37]) is a computationally fast way of solving general deconvolution problems in mixture models, but it can also be used to compute density estimates.

For a general kernel mixture, Newton's algorithm may be described as follows: Assume that $Y_1, \ldots, Y_m \overset{iid}{\sim} h(y) = \int k(y;\theta)\psi(\theta)d\nu(\theta)$, where the mixture density $\psi(\theta)$ with respect to the dominating measure $\nu(\theta)$ is to be estimated. Start with an initial estimate $\psi_0(\theta)$, such as the prior mean, of $\psi(\theta)$. Fix weights $1 \ge w_1 \ge w_2 \ge \cdots w_m > 0$ such as $w_i = i^{-1}$. Recursively

compute

$$\psi_i(\theta) = (1 - w_i)\psi_{i-1}(\theta) + w_i \frac{k(Y_i;\theta)\psi_{i-1}(\theta)}{\int k(Y_i;t)\psi_{i-1}(t)d\nu(t)}, \quad i = 2, \ldots, m,$$

and declare $\psi_m(\theta)$ as the final estimate $\hat{\psi}(\theta)$. The estimate is not a Bayes estimate (it depends on the ordering of the observations), but it closely mimics the Bayes estimate with respect to a DPM prior with kernel $k(x;\theta)$ and center measure with density $\psi_0(\theta)$. If $\sum_{i=1}^{\infty} w_i = \infty$ and $\sum_{i=1}^{\infty} w_i^2 < \infty$ then the mixing density is consistently estimated ([37], [28]).

In the multiple testing context, ν is the sum of point mass at 0 of size 1 and the Lebesgue measure on $(0, 1)$. Then π is identified as $\psi(0)$ and $F(\alpha)$ as $\psi(0) + \int_{(0,\alpha]} \psi(\theta)d\theta$. Then a reasonable estimate is obtained by $\hat{\psi}(0)\alpha/[\hat{\psi}(0)\alpha + \int_{(0,\alpha]} \hat{\psi}(\theta)d\theta]$. The computation is extremely fast and the performance of the estimator is often comparable to that of the Bayes estimator.

Since π takes the most important role in the expression for the pFDR function, it is important to estimate π consistently. However, a conceptual problem arises because π is not uniquely identifiable from the mixture representation $F(x) = \pi x + (1 - \pi)H(x)$, where $H(\cdot)$ is another c.d.f. on [0,1]. Note that the class of such distributions is weakly closed. The components π and H can be identified by imposing the additional condition that H cannot be represented as a mixture with another uniform component, which, for the case when H has a continuous density h, translates into $h(1) = 0$. Letting $\pi(F)$ be the largest possible value of π in the representation, it follows that $\pi(F)$ upper bounds the actual proportion of null hypothesis and hence the actual pFDR is bounded by $\overline{\text{pFDR}}(F;\alpha) := \pi(F)\alpha/F(\alpha)$. This serves the purpose from a conservative point of view. The functional $\pi(F)$ and the $\overline{\text{pFDR}}$ are upper semicontinuous with respect to the weak topology in the sense that if $F_n \to_w F$, then $\limsup_{n\to\infty} \pi(F_n) \leq \pi(F)$ and $\limsup_{n\to\infty} \overline{\text{pFDR}}(F_n;\alpha) \leq \overline{\text{pFDR}}(F;\alpha)$.

Full identifiability of the components π and H in the mixture representation is possible under further restriction on F if $H(x)$ has a continuous density h with $h(1) = 0$ or the tail of H at 1 is bounded by $C(1 - x)^{1+\epsilon}$ for some $C, \epsilon > 0$. The second option is particularly attractive since it also yields continuity of the map taking F to π under the weak topology. Thus posterior consistency of estimating F under the weak topology in this case will imply consistency of estimating π and the pFDR function, uniformly on compact subsets of $(0, 1]$. The class of distributions satisfying the lat-

ter condition will be called $\mathcal{B}$ and $\mathcal{D}$ will stand for the class of continuous decreasing densities on $(0, 1]$.

Consider a prior Π for H supported in $\mathcal{B} \cap \mathcal{D}$ and independently a prior μ for π with full support on $[0, 1]$. Let the true value of π and h be respectively π_0 and h_0 where $0 < \pi_0 < 1$ and $H_0 \in \mathcal{B} \cap \mathcal{D}$. In order to show posterior consistency under the weak topology, we apply Schwartz's result [55]. Clearly we need the true p-value density to be in the support of the beta mixture prior. A density h happens to be a pointwise mixture of be(a, b) with $a < 1$ and $b \geq 1$ if $H(e^{-y})$ or $1 - H(1 - e^{-y})$ is completely monotone, that is, has all derivatives which are negative for odd orders and positive for even orders. Since pointwise approximation is stronger than L_1-approximation by Scheffe's theorem, densities pointwise approximated by beta densities are in the L_1-support of the prior in the sense that $\Pi(h : \|h - h_0\|_1 < \epsilon) > 0$ for all $\epsilon > 0$. Because both the true and the random mixture densities contain a uniform component, both densities are bounded below. Then a relatively simple analysis shows that the Kullback–Leibler divergence is essentially bounded by the L_1-distance up to a logarithmic term, and hence $f_0 = \pi_0 + (1 - \pi_0)h_0$ is in the Kullback–Leibler support of the prior on $f = \pi + (1 - \pi)h$ induced by Π and μ. Thus by the consistency result discussed in Section 7.1 applies so that the posterior for F is consistent under the weak topology. Hence under the tail restriction on H described above, posterior consistency for π and pFDR follows. Even if the tail restriction does not hold, a one-sided form of consistency, which may be called "upper semi-consistency", holds: For any $\epsilon > 0$, $\Pr(\pi < \pi_0 + \epsilon | X_1, \ldots, X_m) \to 1$ a.s. and that the posterior mean $\hat{\pi}_m$ satisfies $\limsup_{m\to\infty} \hat{\pi}_m \leq \pi_0$ a.s.

Unfortunately, the latter has limited significance since typically one would not like to underestimate the true π_0 (and the pFDR) while overestimation is less serious. When the beta mixture prior is used on h with the center measure of the Dirichlet process G_0 supported in $(0, 1) \times (1 + \epsilon, \infty)$ and h_0 is in the L_1-support of the Dirichlet mixture prior, then full posterior consistency for estimating π and pFDR holds. Since the Kullback–Leibler property is preserved under mixtures by Fubini's theorem, the result continues to hold even if the precision parameter of the Dirichlet process is obtained from a prior and the center measure G_0 contains hyperparameters.

7.3.2. *Dependent case: Skew-normal mixture model for probit p-values*

Due to the lack of a suitable multivariate model for the joint distribution of the p-values, most applications assume that the data associated with the family of tests are independent. However, empirical evidence obtained in many important applications such as fMRI, proteomics (two-dimensional gel electrophoresis, mass-spectroscopy) and microarray analysis, shows that the data associated with the different tests for multiple hypotheses are more likely to be dependent. In an fMRI example, tests regarding the activation of different voxels are spatially correlated. In diffusion tensor imaging problems, the diffusion directions are correlated and generate dependent observations over a spatial grid. Hence, a grid-by-grid comparison of such images across patient groups will generate several p-values that are highly dependent.

The p-values, X_i, take values in the unit interval on which it is hard to formulate a flexible multivariate model. It is advantageous to transform X_i to a real-valued random variable Y_i, through a strictly increasing smooth mapping $\Psi : [0,1] \to \mathbb{R}$. A natural choice for Ψ is the probit link function, Φ^{-1}, the quantile function of the standard normal distribution. Let $Y_i = \Phi^{-1}(X_i)$ be referred to as the *probit p-values*. We shall build flexible nonparametric mixture models for the joint density of $(Y_1, \ldots, Y_m)$.

The most obvious choice of a kernel is an m-variate normal density. Efron (cf. [17]) advocated in favor of this kernel. This can automatically include the null component, which is the standard normal density after the probit transformation of the uniform. However, the normal mixture has a shortcoming. As in the previous subsection, marginal density of a p-value is often decreasing. Thus the model on the probit p-values should conform to this restriction whenever it is desired so. The transformed version of a normal mixture is not decreasing for any choice of the mixing distribution unless all components have variance exactly equal to one. This prompts for a generalization of the normal kernel which still includes the standard normal as a special case but can reproduce the decreasing shape of the p-value density by choosing the mixing distribution appropriately. [44] suggested using the multivariate skew-normal kernel as a generalization of the normal kernel. The mixture of skew-normal distribution does provide decreasing p-value densities for a large subset of parameter configurations.

To understand the point, it is useful to look at the unidimensional case. Let

$$q(y; \mu, \omega, \lambda) = 2\phi(y; \mu, \omega^2)\Phi(\lambda\omega^{-1}y)$$

denote the skew-normal density (cf. [2]) with location parameter μ, scale parameter ω and shape parameter λ, where $\phi(y;\mu,\omega^2)$ denotes the $N(\mu,\omega^2)$ density and $\Phi(\cdot)$ denotes the standard normal c.d.f. The skew-normal family has got a lot of recent attention due to its ability to naturally generalize the normal family to incorporate skewness and form a much more flexible class. The skewness of the distribution is controlled by λ and when $\lambda = 0$, it reduces to the normal distribution. If Y has density $q(y;\mu,\omega,\lambda)$, then [44] showed that the density of $X = \Phi(Y)$ is decreasing in $0 < x < 1$ if and only if

$$\omega^2 \geq 1,\ \lambda > \sqrt{(\omega^2-1)/\omega^2} \text{ and } \mu < \lambda H^*(\beta_1(\omega^2,\lambda)),$$

where $\beta_1(\omega^2,\lambda) = (\omega^2-1)/(\lambda^2\omega^2)$, $0 \leq \beta_1 \leq 1$, $H^*(\beta_1) = \inf_x[H(x) - \beta_1 x]$ and $H(\cdot)$ is the hazard function of the standard normal distribution. Now, since the class of decreasing densities forms a convex set, it follows that the decreasing nature of the density of the original p-value X will be preserved even when a mixture of skew-normal density $q(y;\mu,\omega,\lambda)$ is considered, provided that the mixing measure K is supported on

$$\{(\mu,\omega,\lambda) : \mu \leq m(\beta_1(\omega,\lambda)), \quad \omega \geq 1,\ \lambda \geq \sqrt{(\omega^2-1)/\omega^2}\}.$$

Location-shape mixtures of skew-normal family holding the scale fixed at $\omega = 1$ can be restricted to produce decreasing p-value densities if the location parameter is negative and shape parameter is positive. For scale-shape mixtures with the location parameter set to zero, the induced p-value densities are decreasing if the mixing measure has support on $\{(\omega,\lambda) : \omega \geq 1, \lambda \geq \sqrt{1-\omega^{-2}}\}$. Location-scale mixtures with the shape parameter set to zero is the same as location-scale mixtures of normal family. It is clear from the characterization that the normal density is unable to keep the shape restriction. This is the primary reason why we do not work with normal mixtures.

By varying the location parameter μ and the scale parameter ω in the mixture, we can generate all possible densities. The skew-normal kernel automatically incorporates skewness even before taking mixtures, and hence it is expected to lead to a parsimonious mixture representation in presence of skewness, commonly found in the target density. Therefore we can treat the mixing measure K to be a distribution on μ and ω only and treat λ as a hyperparameter. The nonparametric nature of K can be maintained by putting a prior with large weak support, such as the Dirichlet process. A recent result of [61] shows that nonparametric Bayesian density estimation based on a skew-normal kernel is consistent under the weak topology,

adding a strong justification for the use of this kernel. Interestingly, if the theoretical standard normal null distribution is way off from the empirical one, then one can incorporate this feature in the model by allowing K to assign weights to skew-normal components stochastically larger than the standard normal.

In the multidimensional case, [44] suggested replacing the univariate skew-normal kernel by a multivariate analog. [3] introduced the multivariate skew-normal density

$$SN_m(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Omega},\boldsymbol{\alpha}) \equiv 2\phi_m(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})),$$

where ϕ_m is the m-variate normal density. Somewhat more flexibility in separating skewness and correlation is possible with the version of [45].

[44] considered a scale-shape mixture under restriction to illustrate the capability of the skew-normal mixture model. Most commonly arising probit p-value densities can be well approximated by such mixtures. Analogous analysis is possible with mixtures of location, scale and shape. Consider an $m \times m$ correlation matrix $\boldsymbol{R}$ with possibly a very sparse structure. Let $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_m)^T$, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_m)^T$ and $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_m)^T$. Let H_i denote the indicators that the ith null hypothesis H_{i0} is false and let $\boldsymbol{H} = (H_1,\ldots,H_m)^T$. Then a multivariate mixture model for $\boldsymbol{Y} = (Y_1,\ldots,Y_m)^T$ is $(Y|\boldsymbol{\omega},\boldsymbol{\lambda},\boldsymbol{H},\boldsymbol{R}) \sim SN_m(0;\boldsymbol{\Omega},\boldsymbol{\alpha})$ where $\boldsymbol{\Omega} = \boldsymbol{\Delta}_\omega\boldsymbol{R}\boldsymbol{\Delta}_\omega$, $\boldsymbol{\Delta}_\omega = \mathrm{diag}(\boldsymbol{\omega})$ is the diagonal matrix of scale parameters and $\boldsymbol{\alpha} = \boldsymbol{R}^{-1}\boldsymbol{\lambda}$ is the vector of shape parameters. Let H_i be i.i.d. Bernoulli$(1-\pi)$, and independently

$$(\omega_i,\lambda_i)|\boldsymbol{H} \;\sim\; \begin{cases}\delta_{1,0}, & \text{if } H_i = 0,\\ K_0, & \text{if } H_i = 1.\end{cases}$$

The skew-mixture model is particularly suitable for Bayesian estimation. [44] described an algorithm for obtaining posterior samples. Using a result from [3], one can represent $Y_i = \omega_i\delta_i|U| + \omega_i(1-\delta_i^2)V_i$, where $\delta_i = \lambda_i/(1+\lambda_i^2)$, U is standard normal and $\boldsymbol{V} = (V_1,\ldots,V_n)^T$ is distributed as n-variate normal with zero mean and dispersion matrix $\boldsymbol{R}$ independently of U. This representation naturally lends itself to an iterative MCMC scheme. The posterior sample for the parameters in $\boldsymbol{R}$ can be used to validate the assumption of independence. Also, using the posterior samples it is possible to predict the FDP.

It is not obvious how to formulate an analog of Newton's estimate for dependent observations, but we outline the sketch of a strategy below. If the joint density under the model can be factorized as

$Y_1|\theta \sim k_1(Y_1;\theta)$, $Y_2|(Y_1,\theta) \sim k_2(Y_2;Y_1,\theta)$, ..., $Y_m|(X_1,\ldots,Y_{m-1},\theta) \sim k_m(Y_m;Y_1,\ldots,Y_{m-1},\theta)$, then the most natural extension would be to use

$$\psi_i(\theta) = (1-w_i)\psi_{i-1}(\theta) + w_i \frac{k_i(Y_i;Y_1,\ldots,Y_{i-1}\theta)\psi_{i-1}(\theta)}{\int k_i(Y_i;Y_1,\ldots,Y_{i-1},t)\psi_{i-1}(t)d\nu(t)}. \quad (7.4)$$

Such factorizations are often available if the observations arise sequentially. On the other hand, if m is small and $(Y_i|Y_j, j \neq i)$ are simple, we may use the kernel $k_i(y_i|\theta, y_j, j \neq i)$. More generally, if the observations can be associated with a decomposable graphical model, we can proceed by fixing a perfect order of cliques and then reducing to the above two special cases through the decomposition.

7.4. Areas of Application

Multiple testing procedures have gained increasing popularity in statistical research in view of their wide applicability in biomedical applications. Microarray experiments epitomize the applicability of multiple testing procedures because in microarray we are faced with a severe multiplicity problem where the error rapidly accumulates as one tests for significance over thousands of gene locations. We illustrate this point using a dataset obtained from the National Center for Biotechnology Information (NCBI) database. The data comes from an analysis of isografted kidneys from brain dead donors. Brain death in donors triggers inflammatory events in recipients after kidney transplantation. Inbred male Lewis rats were used in the experiment as both donors and recipients, with the experimental group receiving kidneys from brain dead donors and the control group receiving kidneys from living donors. Gene expression profiles of isografts from brain dead donors and grafts from living donors were compared using a high-density oligonucleotide microarray that contained approximately 25,000 genes. [6] analyzed this dataset using a finite skew-mixture model where the mixing measure is supported on only a finite set of parameter values. Due to the high multiplicity of the experiment, even for a single step procedure with a very small α, the FDR can be quite large. [6] estimated that the pFDR for testing for the difference between brain dead donors and living donors at each of the 25,000 gene locations at a fixed level $\alpha = 0.0075$ is about 0.2. The mixture model framework also naturally provides estimates of effect size among the false null. While [6] looked at one sided t-test at each location to generate the p-values, they constructed the histogram of the 25,000 p-values generated from two-sided tests. The

left panel of Figure 7.1 gives a default MATLAB kernel-smoother estimate of the observed p-value histogram. The density shows a general decreasing shape except for local variation and at the edges. The spikes at the edges are artifacts of the smoothing mechanism. The bumpy nature of the smoothed histogram motivates a mixture approach to modeling. The histogram in the probit scale is shown as the jagged line in the right panel in Figure 7.1. The smoothed curve is an estimate of the probit p-value density based on a skew-normal mixture model. Empirical investigation reveals the possibility of correlation among the gene locations. Thus, the multivariate skew-normal mixture would yield more realistic results by incorporating flexible dependence structure.

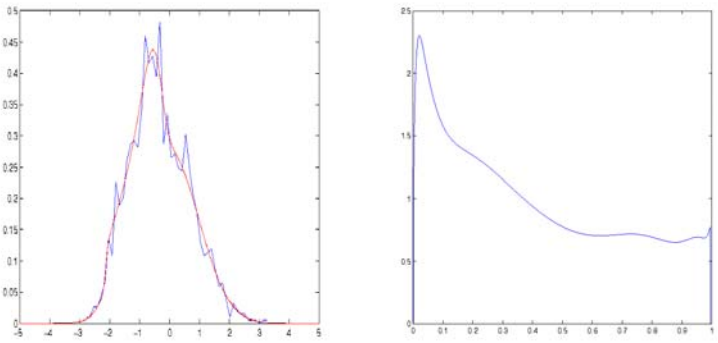

Fig. 7.1. Density of p-values obtained from the ratdata: original scale (left) and probit scale (right).

Another important application area of the FDR control procedure is fMRI. In fMRI data, one is interested in testing for brain activation in thousands of brain voxels simultaneously. In a typical experiment designed to determine the effect of covariate (say a drug or a disease status) on brain activation during a specific task (say eye movement), the available subjects will be divided into the treatment group (individual taking the drug or having a particular disease) and the control group (individuals taking a placebo or not having a disease) and their brain activation (blood oxygen level dependent signal) will be recorded at each voxel in a three dimensional grid in the brain. Then for each of the thousands of voxels, the responses for the individuals in both groups are recorded and then two sample tests are carried out voxel-by-voxel to determine the voxels with significant signal difference

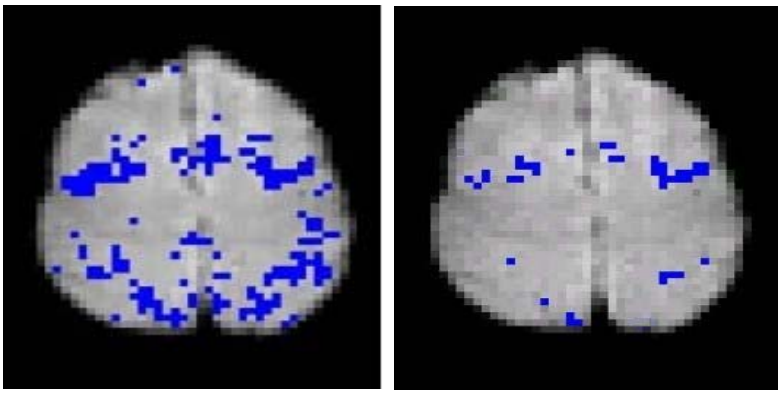

Fig. 7.2. fMRI slice activation image before and after FDR control.

across groups. However due to severe multiplicity, too many voxels may be declared as significant discoveries. Many of these voxels can be adjudged unimportant based on physiological knowledge, but still many others may remain as potential discoveries. The left panel of Figure 7.2 shows the voxels discovered as significant in a particular slice of the brain in a typical fMRI study (the details of the study are not given due to confidentiality issues, the figure is just used for illustration). The stand alone voxels with differential activation are potentially false discoveries where the contiguous clusters of voxels with significant activation pattern are potentially more meaningful findings. However, one needs to use statistical procedures to determine this as there will be tens of thousands of voxels and determining the validity of the findings manually is an infeasible task and a source of potential subjective bias. FDR control has been advocated by [23] to control for false discoveries in fMRI experiments. An application of the B-H procedure removes most of the voxels as false discoveries while keeping only a few with strong signal difference among the two groups. Thus the B-H procedure for this application turns out to be very conservative, and conflicts with scientific goal of finding anatomically rich activation patterns. An FDR control procedure that takes the dependence among voxels into account will be be more appropriate for this application. Work is underway to evaluate the merits of the dependent skew-mixture procedure in a typical fMRI dataset.

[6] also gave an illustration of the pitfalls of constraining the p-value model to have a theoretical null component. In their example, the null com-

ponents were made up of two components, one which is slightly stochastically smaller than the theoretical null and the other which is slightly bigger. With a single theoretical null distribution fitted to the data, both components were poorly estimated while the unconstrained fit with no prespecified theoretical null distribution gave an adequate approximation of both components.

Of course the applicability of multiple testing procedures is not restricted to biomedical problems. While the biomedical problems have been the primary motivation for developing false discovery control procedures, FDR control procedures are equally important in other fields, such as astronomy, where one may be interested in testing significance of findings of several celestial bodies simultaneously. There are important applications in reliability, meteorology and other disciplines as well.

References

[1] Antoniak, C. (1974). Mixtures of Dirichlet processes with application to Bayesian non-parametric problems. *Ann. Statist.* **2** 1152–1174.

[2] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12** 171–178.

[3] Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83** 715–726.

[4] Bayarri, M. J. and Berger, J. O. (2000). p-values for composite null models. *J. Amer. Statist. Asooc.* **95** 1127–1142.

[5] Bazan, J. L., Branco, M. D. and Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis* **1** 861–892.

[6] Bean, G. J., DeRose, E. A., Mercer, L. D., Thayer, L. K. and Roy, A. (2008). Finite skew-mixture models for estimation of false discovery rates. Preprint.

[7] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300.

[8] Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Behav. Educ. Statist.* **25** 60–83.

[9] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188.

[10] Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.* **100** 71–93.

[11] Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika* **93** 491–507.

[12] Black, M. A. (2004). A note on adaptive control of false discovery rates. *J. R. Statist. Soc. Ser. B* **66** 297–304.

[13] Chen, J. and Sarkar, S. K. (2004). Multiple testing of response rates with a control: A Bayesian stepwise approach. *J. Statist. Plann. Inf.* **125** 3–16.
[14] Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion. In *Bayesian Statistics* **2** (J. M. Bernardo, et al., eds.) North-Holland, Amsterdam, 133–156.
[15] Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160.
[16] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104.
[17] Efron, B. (2005). Local false discovery rates. Available at http://www-stat.stanford.edu/ brad/papers/False.pdf.
[18] Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588.
[19] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.
[20] Ferguson, T. S. (1983). Bayesian density estimation by mixtures of Normal distributions. In *Recent Advances in Statistics* (Rizvi M., Rustagi, J. and Siegmund, D., Eds.) 287–302.
[21] Finner, H. and Roters, M. (2002). Multiple hypothesis testing and expected number of type I errors. *Ann. Statist.* **30** 220–238.
[22] Gavrilov, Y., Benjamini, Y. and Sarkar, S. K. (2008). An adaptive step-down procedure with proven FDR control. *Ann. Statist.* (in press).
[23] Genovese, C. R., Lazar, N. A. and Nichols, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* **15** 870–878.
[24] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B* **64** 499–517.
[25] Genovese, C. R. and Wasserman, L. (2004). A stochastic process approach to false discovery rate. *Ann. Statist.* **32** 1035–1063.
[26] Genovese, C. R., Roeder, K. and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524.
[27] Ghosal, S., Roy, A. and Tang, Y. (2008). Posterior consistency of Dirichlet mixtures of beta densities in estimating positive false discovery rates. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (E. Pena *et al.*, eds.), Institute of Mathematical Statistics Collection **1** 105–115.
[28] Ghosh, J. K. and Tokdar, S. T. (2006). Convergence and consistency of Newton's algorithm for Estimating Mixing Distribution. *The Frontiers in Statistics* (J. Fan and H. Koul, eds.), Imperial College Press.
[29] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–802.
[30] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6** 65–70.

[31] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modfied Bonferroni test. *Biometrika* **75** 383–386.
[32] Lehmann, E. L. and Romano, J. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33** 1138–1154.
[33] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *Ann. Statist.* **12** 351–357.
[34] MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7** 223–228.
[35] Miller, R. G. Jr. (1966). *Simultaneous Statistical Inference*. McGraw Hill, New York.
[36] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265.
[37] Newton, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhya Ser. A* **64** 1–17.
[38] Newton, M. A., Quintana, F. A. and Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey *et al.*, eds.) 45–61. Springer-Verlag, New York.
[39] Newton, M. A. and Zhang, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika* **86** 15–26.
[40] Parker, R. A. and Rothenberg, R. B. (1988). Identifying important results from multiple statistical tests. *Statistics in Medicine* **7** 1031–1043.
[41] Robins, J. M., van der Vaart, A. W. and Ventura, V. (2000). Asymptotic distribution of p-values in composite null models. *J. Amer. Statist. Asooc.* **95** 1143–1167.
[42] Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77** 663–665.
[43] Romano, J. P. and Shaikh, A. M. (2006). Step-up procedures for control of generalizations of the familywise error rate. *Ann. Statist.* **34** (to appear).
[44] Roy, A. and Ghosal, S. (2008). Estimating false discovery rate under dependence; a mixture model approch. Preprint.
[45] Sahu, S. K., Dey, D. K. and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canad. J. Statist.* **31** 129–150.
[46] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30** 239–257
[47] Sarkar, S. K. (2006). False discovery and false non-discovery rates in single-step multiple testing procedures. *Ann. Statist*, **34** 394–415.
[48] Sarkar, S. K. (2007). Step-up procedures controlling generalized FWER and generalized FDR. *Ann. Statist.* **35** 2405–2420.
[49] Sarkar, S. K. (2008). Generalizing Simes' test and Hochberg's step-up procedure. *Ann. Statist.* **36** 337–363.
[50] Sarkar, S. K. (2008).Two-stage step-up procedures controlling FDR. *J. Statist. Plann. Inf.* **138** 1072–1084.
[51] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4** 639–650.

[52] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754.
[53] Saville, D. J. (1990). Multiple comparison procedures: the practical solution. *American Statistician* **44** 174–180.
[54] Soric, B. (1989). Statistical "discoveries" and effect size estimation. *J. Amer. Statist. Assoc.* **84** 608–610.
[55] Schwartz, L. (1965). On Bayes Procedures. *Z. Wahrsch. Verw. Gebiete*, **4** 10–26.
[56] Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many test simultaneously. *Biometrika* **69** 493–502.
[57] Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc., Ser. B* **64** 479–498.
[58] Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035.
[59] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Statist. Soc. Ser. B* **66** 187–205.
[60] Tang, Y., Ghosal, S. and Roy, A. (2007). Nonparametric Bayesian estimation of positive false discovery rates. *Biometrics* **63** 1126–1134.
[61] Wu, Y. and Ghosal, S. (2008). Kullback–Leibler property of general kernel mixtures in Bayesian density estimation. *Electronic J. Statist.* **2** 298–331.