Predicting False Discovery Proportion Under Dependence

Subhashis GHOSAL and Anindya ROY

We present a flexible framework for predicting error measures in multiple testing situations under dependence. Our approach is based on modeling the distribution of the probit transform of the p-values by mixtures of multivariate skew-normal distributions. The model can incorporate dependence among p-values and it allows for shape restrictions on the p-value density. A nonparametric Bayesian scheme for estimating the components of the mixture model is outlined and Markov chain Monte Carlo algorithms are developed. These lead to the prediction of false discovery proportion and related credible bands. An expression for the positive false discovery rate for dependent observations is also derived. The power of the mixture model in estimation of key quantities in multiple testing is illustrated by a simulation study. A dataset on kidney transplant is also analyzed using the methods developed.

KEY WORDS: Dirichlet process mixture; False discovery rate; p-value distribution; Shape restriction; Skew-normal distribution.

1. INTRODUCTION

Multiple testing procedures (MTPs) are among the most important statistical tools for modern biomedical and bioinformatics applications such as DNA microarray analysis, proteomics, and functional magnetic resonance imaging (fMRI). In these typical applications, thousands of hypotheses need to be tested simultaneously and MTPs are employed to control the number of false positive findings. A popular measure of false positives that can be controlled is called the false discovery rate (FDR), introduced by Benjamini and Hochberg (1995). The FDR is defined as $\mathrm{E}(V/\max(R,1))$, where R stands for the total number of rejections and V stands for the number of rejections of true null hypotheses. Benjamini and Hochberg (1995) proposed a multi-step testing procedure that controls the FDR when the test statistics for different hypotheses testing are independent. Storey (2002, 2003) introduced an alternative measure called the positive false discovery rate (pFDR), defined by $\mathrm{E}(V/R|R>0) = \mathrm{FDR}/\mathrm{P}(V>0)$. Since then Sarkar (2002, 2004, 2006, 2007), Genovese and Wasserman (2002, 2004), and Efron (2004, 2007), among others, suggested various modifications to FDR and also to FDR controlling MTPs. An objective Bayesian method for normal experiments was developed by Scott and Berger (2005). A wide range of research focused on the FDR and its control in various applications; for example, see the works of Genovese, Lazar, and Nichols (2002) in neuroimaging; Golub et al. (1999) in cancer research; Efron et al. (2001) in DNA microarrays; Miller et al. (2001) and Hopkins et al. (2002) in astrophysics.

However, one of the biggest challenges that has emerged in the multiple testing literature is to find a way to accommodate dependence among the hypotheses. In most present-day applications like fMRI, proteomics (2D-gel electrophoresis, mass-spectroscopy), and DNA microarray analysis, data show clear evidence of dependence among the hypotheses. In the multiple testing problems associated with these applications, a very large number, possibly in tens of thousands, of correlated p-values are generated. In such data structures, error control procedures based on the working hypothesis of independence may be invalid. Benjamini and Yekutieli (2001), Storey and Tibshirani (2003), Finner, Dickhaus, and Roters (2007), Farcomeni (2007), and Sarkar (2007) showed that some controlling procedures or modifications of them are still able to control the targeted measure of error under certain dependence scenarios. Clarke and Hall (2009) showed that the error control procedures exhibit some robustness properties to dependence assumption, provided that the number of hypotheses is large and the p-value distribution has lighter tail. Nevertheless, FDR controlling procedures designed for independent data tend to be conservative and hence lose power under dependence.

The main objective of this article is to describe a black-box model for incorporating dependence among the hypotheses and subsequently propose an MTP. The model will be based on flexible mixture models for the distribution of p-values and we shall use a nonparametric Bayesian scheme to estimate and predict relevant quantities from the model. The Bayesian approach has the attractive feature of having a posterior probability attached to any given null hypothesis. We show that the general formula for FDR under dependence is too complicated to implement in an MTP. We instead predict the false discovery proportion (FDP) given by $V/\max(R,1)$, using the posterior probabilities obtained from the Bayesian scheme, and build the MTP using the predicted FDP curve. Pawitan, Calza, and Alexander (2006) showed that accuracy of FDP prediction can really suffer if dependence is ignored.

The organization of the article is as follows. In Section 2, we describe a mixture model framework for dependent p-values and discuss key issues from a modeling perspective. Prior specification and posterior computation techniques are described in Section 3. A simulation study and analysis of a real dataset are presented in Section 4. Proofs are presented in the Appendix.

Subhashis Ghosal is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: *sghosal@stat.ncsu.edu*). Anindya Roy is Professor, Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250 (E-mail: *anindya@umbc.edu*). Subhashis Ghosal's research was supported in part by NSF grant DMS-0803540 and Anindya Roy's research was supported in part by NSF grant DMS-0803531. The authors thank the associate editor and two anonymous referees for their detailed comments and suggestions.

Journal of the American Statistical Association
September 2011, Vol. 106, No. 495, Theory and Methods
DOI: 10.1198/jasa.2011.tm10488

2. MODELING DEPENDENT p-VALUES

Our primary objective is to build a flexible model for the p-values that can incorporate dependence among the hypotheses. Before we describe our model, we discuss some of the key modeling issues involved in MTPs.

Consider the problem of testing m simultaneous hypotheses H_{0i} versus H_{1i}, $i = 1, \ldots, m$. Let H_i stand for the indicator that H_{0i} is false. Let $X_1, \ldots, X_m$ denote the associated p-values. Let $\mathbb{1}$ denote the indicator function and $I_i = I_i(\gamma) = \mathbb{1}\{X_i < \gamma\}$ denote the indicator that H_{0i} is rejected at nominal level γ, $i = 1, \ldots, m$, and $\mathbf{I} = (I_1, \ldots, I_m)$. Then the number of rejected hypotheses is $R = \sum_{i=1}^m I_i$ and the number of falsely rejected hypothesis is $V = \sum_{i=1}^m I_i(1 - H_i)$.

Assuming that H_i are iid with $\mathrm{P}(H_i = 0) = 1 - \mathrm{P}(H_i = 1) = \pi_0$, $i = 1, \ldots, m$, Storey (2002) showed that the pFDR is given by the expression $\pi_0\gamma/F(\gamma)$, when all hypotheses have a common nominal significance level γ and F denotes the marginal distribution of the p-values. The corresponding expression for FDR is $\{\pi_0\gamma/F(\gamma)\}\mathrm{P}(V > 0)$. In Storey's approach, once π_0 and F are estimated, the obtained pFDR can be used to estimate the nominal level γ for which the pFDR is controlled at a desired level. Under this setup, the marginal p-value density is $f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_1(x)$, where $f_0(x)$ is the p-value density under the null and $f_1(x)$ is that under the alternative.

2.1 FDR Under Dependence

Unfortunately, under dependence the simple expression for pFDR is no longer valid. The following result derives the expression for FDR (and hence pFDR) under dependence. The FDP at γ is given by

$$\mathrm{FDP}(\gamma) = \frac{V}{\max(R, 1)} = \frac{\sum_{i=1}^m I_i(1 - H_i)}{\sum_{i=1}^m I_i + \prod_{i=1}^m (1 - I_i)}. \tag{1}$$

By definition, $\mathrm{FDR}(\gamma) = \mathrm{E}(\mathrm{FDP}(\gamma))$ and $\mathrm{pFDR}(\gamma) = \mathrm{FDR}(\gamma)/\mathrm{P}(V > 0)$. Let $\mathcal{B}_i^m$ denote the set of all m-dimensional binary vectors with a 1 at the ith position.

Theorem 1. For arbitrarily dependent observations,

$$\mathrm{FDR}(\gamma) = \pi_0 \sum_{i=1}^m b_i(\gamma), \tag{2}$$

where

$$b_i(\gamma) = \sum_{\mathbf{a} \in \mathcal{B}_i^m} \frac{\mathrm{P}(\mathbf{I} = \mathbf{a} | H_i = 0)}{\sum_{i=1}^m a_i}.$$

If the observations are exchangeable, then

$$\mathrm{FDR}(\gamma) = \pi_0 m b_1(\gamma). \tag{3}$$

When the observations are independent and identically distributed, (3) reduces to the familiar expression $\mathrm{FDR}(\gamma) = \frac{\pi_0\gamma}{F(\gamma)}\mathrm{P}(R > 0)$.

Even for the exchangeable case, numerical evaluation of $\mathrm{FDR}(\gamma)$ can be challenging. This puts a major hurdle in dealing with the FDR in the dependent case. As argued later, a more prudent strategy from a computational point of view is to concentrate on predicting the FDP process and use the predicted FDP to choose a nominal level for the hypotheses. Thus, we seek a suitable model for the dependent p-value under which we can effectively predict the FDP process.

A model for dependent p-values (and hence dependent hypotheses) has to be flexible enough to accommodate a wide range of testing problems. Such a model has primarily two components: a flexible model for $f(x)$ which can incorporate several known features of the marginal density and a description of a family of multivariate densities with desired correlation structure and having marginal density $f(x)$.

Unfortunately, standard univariate p-value densities do not have flexible multivariate generalizations due to the restriction on the range of values. To model the joint behavior of p-values, it is easier to work with a transformation that removes the restriction. We choose the probit transformation $Y_i = \Phi^{-1}(X_i)$ of the p-values to develop the multivariate model. We consider mixtures of multivariate kernels as possible model for the probit p-values because they provide sufficient flexibility in modeling. The necessity for considering mixtures can be illustrated through the following example. Consider testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$ based on an observation U following a Cauchy density with location parameter θ. Suppose a test rejects H_0 for large values of U. Then the p-value, X, is given by $\cot(\pi U)$, and hence its density under an alternative $\theta > 0$ is given by $\{1 - \theta \sin(2\pi x) + \theta^2 \sin^2(\pi x)\}^{-1}$ for $0 < x < 1$. The density of the probit transform $Y = \Phi^{-1}(X)$ is plotted in Figure 1(a), which shows existence of bumps that typically cannot be captured by a single parametric density. The versatility of mixtures can easily accommodate features such as bumps and shoulders appearing in the probit p-value density.

In most testing problems, the null density is uniform. This holds, for instance, if H_{0i} is simple, or reduced to a simple hypothesis by similarity or invariance, or holds true approximately for general composite hypothesis if partial predictive p-value is considered; see the work of Bayarri and Berger (2000) and Robins, van der Vaart, and Ventura (2000). Under the probit transformation the null density becomes standard normal. Thus, the mixture family should include normal components to model the null p-value density. Efron (2004) cautioned that in some situations, the overall null distribution of all p-values may deviate from the uniform, in which case the empirical null distribution should be considered.

The mixture model for the joint densities of the probit p-values can be built by first conditioning on the latent indicators $H_1, \ldots, H_m$. Consider a multivariate family of densities $p(\mathbf{y}; \theta_1, \ldots, \theta_m, \rho)$ which includes the family of multivariate normal with means zero and variances 1 and some correlation structure driven by the additional parameter ρ. Let θ_0 stand for a null value of θ which gives rise to $\mathrm{N}(0, 1)$ marginal distribution for the corresponding component and let $\delta(\cdot; \theta_0)$ stand for the probability measure degenerate at θ_0. Model the joint density of $\mathbf{Y} = (Y_1, \ldots, Y_m)$ given $\mathbf{H} = (H_1, \ldots, H_m)$ by the following hierarchical scheme:

$$\begin{aligned} \mathbf{Y} | \mathbf{H} &\sim p(\mathbf{y}; \theta_1, \ldots, \theta_m, \rho), \\ \theta_i &\sim \begin{cases} \delta(\cdot; \theta_0), & \text{if } H_i = 0, \\ G, & \text{if } H_i = 1. \end{cases} \end{aligned} \tag{4}$$

It appears that a multivariate normal family with general mean, variances, and a given correlation structure will be a very

natural candidate for the family of kernels $p(\mathbf{y}; \theta_1, \ldots, \theta_m, \psi)$. Unfortunately, if normal mixtures are used for probit p-values, they fail to impose one of the known restrictions on the shape of the marginal p-value density under the alternative, namely decreasing density under the alternative. The marginal p-value density can be naturally written as the mixture $f(y) = \pi_0 + (1 - \pi_0) f_1(y)$ where f_1 is the density under the alternative. The corresponding mixture form for the probit p-value density would be $h(y) = \pi_0 \phi(y) + (1 - \pi_0) h_1(y)$, where $\phi(\cdot)$ is the standard normal density. Apart from a known null density, there are additional features in the alternative density $f_1(x)$ that can provide increase in efficiency of MTPs if those are properly accounted for. In general, the alternatives are composite hypotheses and hence the p-value density under the alternative is generally a mixture whose components are p-value densities associated with specific alternatives. The density under the alternative is concentrated near zero. Moreover, under natural conditions like monotone likelihood ratio property the alternative density $f_1(x)$ is decreasing (cf. propositions 1 and 2 of Ghosal, Roy, and Tang 2008). A natural mixture kernel on $[0, 1]$ that provides the decreasing shape restriction is the Beta(a, b) density with $a < 1$, $b \geq 1$. Tang, Ghosal, and Roy (2007) used the mixture model

$$f(x) = \pi_0 + (1 - \pi_0) \int_{(0,1)\times[1,\infty)} (\Gamma(a)\Gamma(b)/\Gamma(a+b)) \times x^{a-1}(1-x)^{b-1}\, dG(a, b),$$

where G is a mixing measure.

We consider a broader family of skew-normal mixtures in our modeling. The skew-normal distribution was introduced by Azzalini (1985) and generalized by Azzalini and Dalla Valle (1996) in the multivariate situation as a flexible yet very tractable generalization of the normal family that incorporates skewness into consideration and has found a wide range of applications; see the book edited by Genton (2004) for details. To illustrate the ideas behind modeling by skew-normal mixtures, we begin with the univariate case and generalize to the multivariate setting afterward.

Let $q(y; \mu, \omega, \lambda)$ denote the univariate skew-normal density with location parameter μ, scale parameter ω, and shape parameter λ given by

$$q(y; \mu, \omega, \lambda) = 2\phi(y; \mu, \omega)\Phi(-\lambda\omega^{-1}(y - \mu)), \qquad (5)$$

where $\phi(y; \mu, \omega)$ denotes the $\mathrm{N}(\mu, \omega^2)$ density and the $\Phi(\cdot)$ denotes the standard normal cumulative distribution function (cdf). Then the density of probit p-values under the alternative is modeled as

$$h_1(y) = \int q(y; \mu, \omega, \lambda)\, dG(\mu, \omega, \lambda), \qquad (6)$$

where G is the mixing measure on the space of (μ, ω, λ).

Theorem 2. Suppose that Y has density $q(y; \mu, \omega, \lambda)$ and $X = \Phi(Y)$. Then the density $f_1(x)$ of X is decreasing in $0 \leq x \leq 1$ if and only if

$$\mu \leq \lambda\omega^{-1}\varphi(\lambda^{-2}(\omega^2 - 1)), \qquad \omega \geq 1, \lambda \geq \sqrt{\omega^2 - 1}, \qquad (7)$$

where $\varphi(\beta) = \inf\{H_\Phi(x) - \beta x : x \in \mathbb{R}\}$, and $H_\Phi(x) = \phi(x)/(1 - \Phi(x))$ is the hazard function of the standard normal distribution.

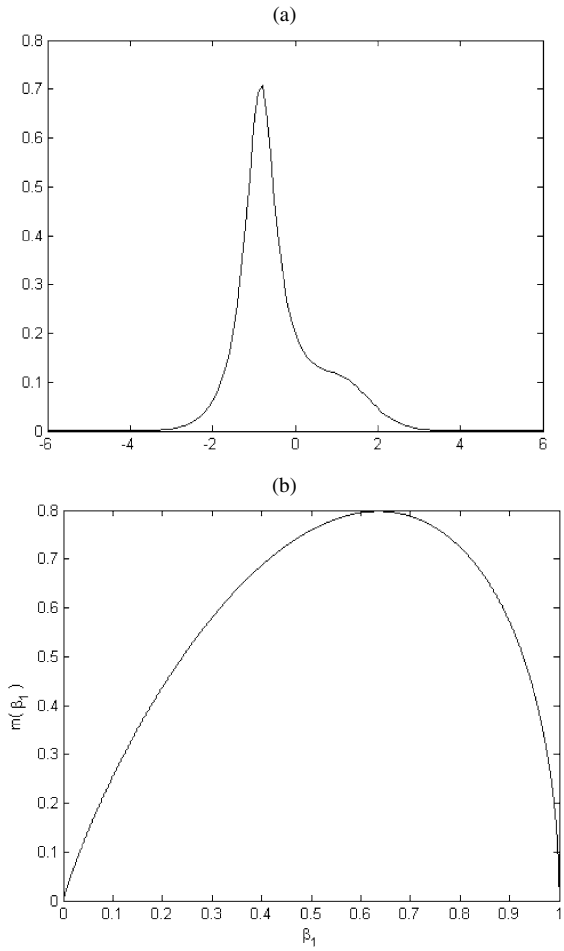

Figure 1. (a) Density of probit p-value for the Cauchy location testing problem; (b) plot of the function $\varphi(\beta)$ described in Theorem 2.

A plot of φ is shown in Figure 1(b). Observe that, as the class of decreasing densities is convex, the skew-normal mixtures produce decreasing densities as long as the mixing distribution is supported on the region defined by (7).

Therefore, the primary reasons for choosing the skew-normal family are the following:

(i) The skew-normal is a broader family than the normal, and hence in particular includes $\mathrm{N}(0, 1)$, which is to be used as the null distribution for probit p-values.

(ii) The joint distribution can model a broad class of correlation structures.

(iii) Individual components of the mixture exhibit appropriate skewness. Even though unrestricted mixtures of symmetric densities with a large number of components can capture skewness, only a few components may be needed if they are already skewed.

(iv) Finally, suitably restricted skew-normal mixtures can produce decreasing shape of the original p-value density as shown below.

2.2 Multivariate Mixture Model for Probit p-Values

The main advantage of the skew-normal mixture model (6) is that it can be generalized to incorporate dependence while maintaining the salient features of the marginal p-value density, leading to a model for the joint distribution of the p-values. The

generalization to the dependent case is achieved by replacing the univariate skew-normal kernel $q(y; \mu, \omega, \lambda)$ by the multivariate skew-normal density (Azzalini and Dalla Valle 1996) given by

$$q_m(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{R}) = 2\phi_m(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Omega})\Phi(-\boldsymbol{\alpha}'\mathbf{D}^{-1}(\mathbf{y} - \boldsymbol{\mu})), \quad (8)$$

where $\boldsymbol{\mu} \in \mathbb{R}^m$, $\boldsymbol{\omega}, \boldsymbol{\lambda} \in (0, \infty)^m$, $\mathbf{R}$ is an $m \times m$ positive definite correlation matrix, $\mathbf{D} = \operatorname{diag}(\boldsymbol{\omega})$, $\boldsymbol{\Omega} = \mathbf{R} + \boldsymbol{\lambda}\boldsymbol{\lambda}'$, $\boldsymbol{\alpha} = \boldsymbol{\Omega}^{-1}\mathbf{D}^{-1}\boldsymbol{\lambda}/\sqrt{1 + \boldsymbol{\lambda}'\mathbf{R}^{-1}\boldsymbol{\lambda}}$, and $\phi_m(\cdot; \boldsymbol{\mu}, \boldsymbol{\Omega})$ stands for the density of $\mathrm{N}_m(\boldsymbol{\mu}, \boldsymbol{\Omega})$. Properties of the multivariate skew-normal density were neatly reviewed by Dalla Valle (2004). It may be noted that there are other multivariate generalizations of the skew-normal family such as that by Sahu, Dey, and Branco (2003). An advantage of the latter is that unlike in (8), the correlation and skewness parameters are completely separately treated, and hence it is easier to use for estimation purposes. Nevertheless, structural properties of (8) make it easier to work with in the present context. Many interesting properties of multivariate skew-normal density as well as posterior computational methods in the next section are driven by the following representations (see Dalla Valle 2004):

$$\mathbf{Y} = \boldsymbol{\mu} + |Z_0|\mathbf{D}\boldsymbol{\delta} + \mathbf{D}\boldsymbol{\Delta}\mathbf{Z}, \quad (9)$$

where $\mathbf{Z} = (Z_1, \ldots, Z_m)'$ and $(Z_0, \mathbf{Z})' \sim \mathrm{N}_{m+1}\big(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0 \\ 0 & \mathbf{R} \end{smallmatrix}\right)\big)$, $\boldsymbol{\Delta} = \operatorname{diag}(\sqrt{1 - \delta_1^2}, \ldots, \sqrt{1 - \delta_m^2}) = \operatorname{diag}((1 + \lambda_1^2)^{-1/2}, \ldots, (1 + \lambda_m^2)^{-1/2})$, and $\delta_j = -\lambda_j/\sqrt{1 + \lambda_j^2}$, $j = 1, \ldots, m$.

From (4) and (8), a multivariate skew-normal mixture model for $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ is given by $\mathbf{Y}|(\boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{R}, \mathbf{H}) \sim q_m(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{\lambda})$, where given $\mathbf{H}$, $(\mu_i, \omega_i, \lambda_i)$ is $(0, 1, 0)$ if $H_i = 0$ and is a draw from the mixing distribution G if $H_i = 1$. The multivariate skew-normal mixture model is not identifiable, that is, the mixing distribution G may not be uniquely obtained from the mixture densities, as is common in most mixture models with a scale parameter. Nevertheless, under appropriate restrictions on the support of G such as the shape restriction condition defined by (7), Ghosal and Roy (2011) showed that the point mass π_0 of the mixing distribution at the point $(0, 1, 0)$ is identifiable.

Typically the entries of the correlation matrix $\mathbf{R}$ will be modeled as functions of a low-dimensional parameter ρ. For instance, if $\mathbf{R}$ is the correlation matrix of the AR(1) process with autocorrelation ρ, then ρ is one-dimensional and all elements of $\mathbf{R}$ are functions of ρ. Similarly, $\mathbf{R}$ can be the correlation matrix of the intraclass correlated process with all off-diagonal entries equal to ρ, $0 \le \rho < 1$.

In most applications, the correlation among the probit p-values is known only in a qualitative manner. For example, it may be known that the test statistics have a nearest-neighbor dependence, which translates to similar relation among p-values. The marginal correlation will be a function of the correlation matrix $\mathbf{R}$ and other parameters. For a given value of $(\boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{\lambda})$, the correlation structure of $\mathbf{Y}$ is a rank-1 modification of $\mathbf{R}$ (cf. Azzalini and Dalla Valle 1996) and hence tends to be very similar to $\mathbf{R}$. While the correlation structure in the model may be simplistic, allowing an undetermined parameter as well as taking mixtures together can model relatively complex structures. It is essential to keep $\mathbf{R}$ simple since inversion will be necessary at each MCMC iteration of posterior computation. If the expected value of the rank-1 modification is small compared to $\mathbf{R}$, then even after mixing, the correlation structure will tend to be similar to that of $\mathbf{R}$.

If the p-values are exchangeable, so are the probit p-values. For example, when $\mathbf{R}$ is the correlation matrix of the intraclass process, the intraclass structure is preserved under probit transforms, although the value of the intraclass correlation is likely to change. A similar property holds under stationarity as the following result shows.

Proposition 1. Let the elements of $\mathbf{R}$ be of the form $r(s, t) = r(|s - t|)$ for some stationary autocorrelation function $r(\cdot)$. Then the p-values and probit p-values follow strictly stationary stochastic processes. Moreover, the p-value process is also covariance stationary.

If the observations arise from an AR(1) process, then the probit p-value process is stationary, and hence an AR(1) structure may provide a reasonable fit even though the AR(1) structure is not exactly correct for the probit p-values. In the simulation section, we observe that prediction of FDP is fairly accurate assuming the AR(1) correlation structure, when the data are generated from an AR(1) process.

3. PRIOR SPECIFICATION AND POSTERIOR COMPUTATION

We first specify a prior distribution for the parameters π_0, ρ and the infinite-dimensional parameter G. Let all of them be a priori independently distributed with

$$\begin{aligned} &\pi_0 \sim \mathrm{Beta}(a_0, a_1), \qquad \rho \sim p_\rho(\cdot), \\ &G \sim \mathrm{DP}(M, G_0), \end{aligned} \quad (10)$$

where p_ρ is a density on the parameter space of ρ (like $(-1, 1)$ for AR(1) and $[0, 1)$ for the intraclass correlated process) and $\mathrm{DP}(M, G_0)$ is the Dirichlet process with precision parameter $M > 0$ and center measure G_0. The center measure G_0 is chosen to be supported on the space defined by inequalities (7) to make the marginal p-value density maintain the decreasing shape restriction. Finally, the center measure G_0 of the Dirichlet process is chosen to be of the form $p_\omega(\omega)p_\lambda(\lambda|\omega)p_\mu(\mu|\omega, \lambda)$, where we take $p_\omega(\omega)$ to be $\mathrm{Gamma}(\alpha_\omega, \beta_\omega)$ shifted to the right by 1, $p_\lambda(\lambda|\omega)$ to be $\mathrm{Gamma}(\alpha_\lambda, \beta_\lambda)$ shifted to the right by $\sqrt{\omega^2 - 1}$, and $p_\mu(\mu|\omega, \lambda)$ to be the negative of $\mathrm{Gamma}(\alpha_\mu, \beta_\mu)$ shifted to the right by $\lambda\omega^{-1}\varphi(\lambda^{-2}(\omega^2 - 1))$. In a Dirichlet process mixture model (DPMM), the random measure G can be integrated out from the prior distribution, leading to the joint prior distribution of $\theta_i = (\mu_i, \omega_i, \lambda_i)$'s given by the generalized Polya urn scheme

$$\begin{aligned} &\theta_1 \sim G_0, \\ &\theta_{i+1}|(\theta_1, \ldots, \theta_i) \sim \frac{M}{M + i}G_0 + \sum_{j=1}^{i} \frac{1}{M + i}\delta(\cdot; \theta_j), \qquad i \ge 1. \end{aligned} \quad (11)$$

As a result, the DPMM generates θ_i values, many of which are likely to be tied, giving a very desirable clustering property allowing huge dimension reduction. Let the distinct values be denoted by $\theta_1^*, \ldots, \theta_N^*$ in a typical realization. MCMC algorithms for posterior computation in DPMM were developed by Escobar and West (1995) and others, and are particularly

useful when the center measure G_0 is conjugate to the likelihood of $(Y_1, \ldots, Y_n)$ given the latent variables. In the absence of such a conjugacy property as in the present case, the "no gaps" algorithm of MacEachern and Müller (1998) or similar refinements are needed. While in principle, the "no gaps" algorithm is applicable, the dependence and complexity of the multivariate skew-normal likelihood makes it very challenging to implement. An alternative approximate method developed by Ishwaran and Zarepour (2000) proposed to fix N beforehand to a sufficiently high number and approximate the Dirichlet process by the so-called Dirichlet sieve process given by

$$\begin{aligned} \theta_i &\sim \sum_{j=1}^{N} p_j \delta(\cdot; \theta_j^*), \\ \theta_1^*, \ldots, \theta_N^* &\overset{\text{iid}}{\sim} G_0, \\ (p_1, \ldots, p_N) &\sim \text{Dirichlet}\left(N; \frac{M}{N}, \ldots, \frac{M}{N}\right). \end{aligned} \tag{12}$$

Let $L_i = j$ if $\theta_i = \theta_j^*$. As $N \to \infty$, the Dirichlet sieve process converges to the corresponding Dirichlet process. The main advantage of this approach is that the MCMC problem is reduced to a relatively low and fixed dimensional sampling problem, where only $(L_i : i = 1, \ldots, m)$ and $(\theta_j^*, p_j : j = 1, \ldots, N)$ need to be generated. When the sample size m is large, Ishwaran and Zarepour (2000) recommended taking N to be about $\sqrt{m}$. Strictly speaking, the approximation is justified only in the prior, although it often gives fairly accurate approximation to the full DPMM posterior. An alternative would be to view the Dirichlet sieve process itself as the prior distribution, which shares almost all desirable features of the full Dirichlet process for large N. We follow the Dirichlet sieve process approach in our computation.

We introduce some notation before we describe the steps of posterior computation. Let $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_N^*)$ denote the set of possible alternative values of μ_i's, while the null value is $\mu_0^* = 0$. For $j = 0, 1, \ldots, N$, let $\boldsymbol{\mu}^{i \to j}$ denote the vector $(\mu_1, \ldots, \mu_{i-1}, \mu_j^*, \mu_{i+1}, \ldots, \mu_m)$. Similarly define $\boldsymbol{\omega}^{i \to j}$ and $\boldsymbol{\lambda}^{i \to j}$, where their null values are 1 and 0, respectively. Then in each step of posterior iteration:

(i) For $i = 1, \ldots, m$, H_i are randomly drawn from a Bernoulli distribution with probability $\text{P}(H_i = 1 | \mathbf{Y}, \text{rest})$ given by

$$\begin{aligned} &(1 - \pi_0) q_m(\mathbf{Y}; \boldsymbol{\mu}^{i \to L_i}, \boldsymbol{\omega}^{i \to L_i}, \boldsymbol{\lambda}^{i \to L_i}, R(\rho)) \\ &\quad / \big(\pi_0 q_m(\mathbf{Y}; \boldsymbol{\mu}^{i \to 0}, \boldsymbol{\omega}^{i \to 0}, \boldsymbol{\lambda}^{i \to 0}, R(\rho)) \\ &\quad + (1 - \pi_0) q_m(\mathbf{Y}; \boldsymbol{\mu}^{i \to L_i}, \boldsymbol{\omega}^{i \to L_i}, \boldsymbol{\lambda}^{i \to L_i}, R(\rho))\big). \end{aligned}$$

(ii) For $i = 1, \ldots, m$, the ith component of the configuration vector, L_i, is updated as a random draw from the distribution

$$\text{P}(L_i = j | \mathbf{Y}, \text{rest}) = \frac{p_j q_m(\mathbf{Y}; \boldsymbol{\mu}^{i \to j}, \boldsymbol{\omega}^{i \to j}, \boldsymbol{\lambda}^{i \to j}, R(\rho))}{\sum_{k=1}^{N} p_k q_m(\mathbf{Y}; \boldsymbol{\mu}^{i \to k}, \boldsymbol{\omega}^{i \to k}, \boldsymbol{\lambda}^{i \to k}, R(\rho))}, \quad j = 1, \ldots, N.$$

(iii) For $i = 1, \ldots, m$, the parameters are then updated as $\theta_i = \theta_{L_i}^*$ if $H_i = 1$, and $\theta_i = \theta_0^*$ if $H_i = 0$.

(iv) The assignment probability vector is updated as

$$(p_1, \ldots, p_N) | (\mathbf{Y}, \text{rest}) \sim \text{Dirichlet}\left(N, \frac{M}{N} + c_1, \ldots, \frac{M}{N} + c_N\right),$$

where $c_j = \#\{i : L_i = j \text{ and } H_i = 1\}$.

(v) The null probability π_0 is updated as a draw from $\text{Beta}(a_0 + m_0, a_1 + m - m_0)$ where $m_0 = \sum_{i=1}^{m} (1 - H_i)$.

(vi) The parameter ρ appearing in the correlation structure is updated using a Metropolis–Hastings step

$$p(\rho | \mathbf{Y}, \text{rest}) \propto p_\rho(\rho) q(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{\lambda}, R(\rho)).$$

Finally, a Metropolis–Hastings step will refresh the distinct values $(\theta_1^*, \ldots, \theta_N^*)$.

The posterior sampling algorithm described above can be used to generate a large number K of posterior samples of the parameters $(\pi_0, \theta_1^*, \ldots, \theta_N^*, p_1, \ldots, p_N)$ and hidden labels $(L_1, \ldots, L_n, H_1, \ldots, H_m)$ and hence also posterior samples of any quantity of interest. For instance, the probit p-value density function for each sample is computed as

$$\begin{aligned} h_1^{(r)}(y) = \pi_0^{(r)} \phi(y) + \big(1 - \pi_0^{(r)}\big) \\ \times \frac{M \int q_m(\mathbf{y}; \theta)\, dG_0(\theta) + \sum_{r=1}^{N} q_m(y; \theta_i^{(r)})}{M + m}, \end{aligned}$$

$r = 1, \ldots, K$, and the corresponding posterior mean may be obtained by averaging. Similarly, the posterior samples of p-value density function may be obtained using the relationship between the p-value density and the probit p-value density, while posterior samples of the cdf of probit p-values can be obtained by computing the cdf of a skew-normal distribution numerically using the technique of Bazan, Branco, and Bolfarine (2006). To predict the FDP, one can similarly compute the posterior mean by plugging in posterior samples in the formula (1). Credible intervals are obtained easily using posterior samples.

4. NUMERICAL ILLUSTRATION

4.1 Simulation Study

To evaluate the performance of the proposed method we perform some simulation experiments.

4.1.1 Data Generation Process. We consider the scenario where each observation is an m-dimensional Gaussian data vector and there are two groups containing n_1 and n_2 such observations, respectively. We assume that both groups have the same covariance structure. Specifically, if $\mathbf{Z}_{gj}$ denotes the jth observation in group g, then

$$\mathbf{Z}_{gj} \sim \text{N}_m(\boldsymbol{\mu}_g, \boldsymbol{\Psi}), \qquad g = 1, 2; j = 1, \ldots, n_i,$$

where the dependence structure across the m coordinates is governed by the covariance matrix $\boldsymbol{\Psi}$. The mean vector for the gth group is $\boldsymbol{\mu}_g = (\mu_{g1}, \ldots, \mu_{gm})$, $g = 1, 2$. We consider testing equality of group means across each of the m coordinates. Thus, we have m simultaneous hypotheses $\text{H}_{0i} : \mu_{1i} = \mu_{2i}$. Let X_i denote the p-value generated from the two-sample t-test of the ith hypothesis and let $Y_i = \Phi^{-1}(X_i)$ be the corresponding probit p-value.

Table 1. Empirical performance measures for the skew-mixture model based on replications

		Intraclass			AR(1)		
π_0	ρ	RMSE($\hat{\pi}_0$)	Coverage	RMSPE	RMSE($\hat{\pi}_0$)	Coverage	RMSPE
0.90	0.10	0.0050	0.94	0.0084	0.0127	0.93	0.0440
0.90	0.50	0.0066	0.89	0.0141	0.0256	0.88	0.0821
0.95	0.10	0.0032	0.91	0.0067	0.0095	0.92	0.0333
0.95	0.50	0.0067	0.90	0.0188	0.0306	0.81	0.0844

4.1.2 Parameter and Prior Specification. For the simulation experiment, we choose $n_1 = n_2 = 15$ and $m = 1000$. The mean vector of the first group is zero and the first $m_0 = \lfloor \pi_0 m \rfloor$ coordinates of $\boldsymbol{\mu}_2$ are chosen to be zero, reflecting about π_0 proportion for true null hypotheses. Of the remaining $m_1 = m - m_0$ coordinates of $\boldsymbol{\mu}_2$, 40% are chosen to be 0.5, 50% are chosen to be 1, and 10% are set to 2. The common variance across groups is 1, and hence $\boldsymbol{\Psi}$ is also the correlation matrix.

The elements of the correlation matrix $\boldsymbol{\Psi} = ((\psi_{ij}))$ are parameterized by a single parameter ρ. We choose two different correlation structures; intraclass where $\psi_{ij}(\rho) = \rho$ if $i \neq j$ and 1 if $i = j$; autoregressive of order 1 where $\psi_{ij}(\rho) = \rho^{|i-j|}$.

The prior for the true null proportion π_0 is chosen as Beta(a_0, a_1) where $a_0 = 5$ and $a_1 = 1$. Thus, a priori the expected proportion of null hypotheses is 83.33%. The prior for the correlation parameter is chosen as Uniform$[-1, 1]$. The number of alternative values is $N = 20$ and the number of hypotheses is $m = 1000$. All parameters of the gamma priors for μ, ω, and λ are set to 2. The formula for inverse of the correlation matrix given by Remark A.1 in the Appendix is used in computation.

4.1.3 Summary Output. Table 1 summarizes the performance of the proposed methodology in terms of the root mean squared error (RMSE) values. The RMSPE stands for the integrated root mean squared error in predicting FDP at the nominal level γ and is given by RMSPE $= [\mathrm{E}(\int_0^1 |\widehat{\mathrm{FDP}}(\gamma) - \mathrm{FDP}(\gamma)|^2 \, d\gamma)]^{1/2}$, where $\widehat{\mathrm{FDP}}$ stands for the posterior mean of FDP. The column indicating coverage gives the coverage of the 90% FDP credible interval when $\gamma = 0.05$. Expectations are calculated by replications of the simulation experiment.

We generated 6000 MCMC samples and used the first 1000 for burn-in, thus using 5000 samples to compute posterior means in every replication. For each dataset, the run time is about 3.5 minutes for the MCMC procedure on a Pentium 4 dual core machine.

When the data dependence is not too strong, the proposed method is accurate with respect to all three measures for both types of correlation structures. When the correlation parameter is high ($\rho = 0.5$), true null proportion is still estimated accurately. However, the accuracy of FDP prediction goes down slightly for the AR(1) case. A possible reason for this is that autoregressive structure on p-values does not exactly translate into an autoregressive structure for the probit p-values used in the modeling. For intraclass correlated data, the induced correlation structure on probit p-values is still exchangeable. Nevertheless, the fairly accurate prediction even in the autoregressive case suggests that matching correlation structure exactly is not critical as long as there is enough flexibility left in the model. Figure 2 shows typical FDP curves with their predicted values and confidence bounds when γ is in the interval $[0, 0.5]$. In most cases, the bounds are quite sharp and yield reasonable coverage.

4.2 Effect of Ignoring Dependence

The nonparametric Bayes procedure described above is able to capture dependence among the hypotheses reasonably well. The main advantage is that, by directly accounting for the dependence, the efficiency of the MTP is significantly improved

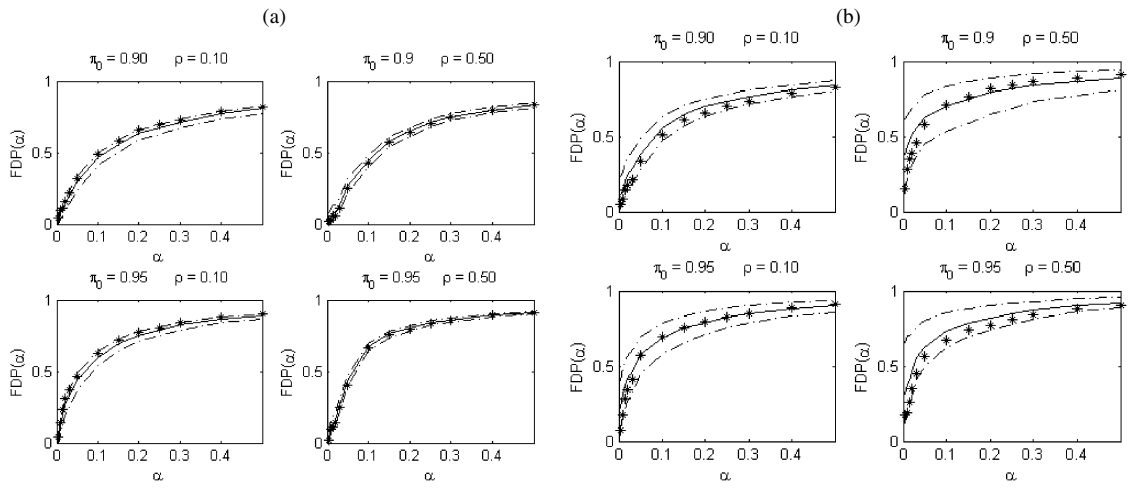

Figure 2. Predicted and true FDP and credible bands for correlation structures: (a) intraclass; (b) AR(1). Solid lines are predicted values with dashed credible limits. The asterisks are the true FDP values.

compared to the procedures that ignore the dependence or those which provide conservative control under dependence. In experiments where moderate to strong dependence is suspected among the test statistics for the hypotheses, the nonparametric Bayesian procedure is expected to improve the power of the procedure while providing desired control on FDR. To evaluate the effect of dependence on the MTPs, we perform a limited simulation study. Our main objective is to compare the FDR obtained from different procedures for certain desired levels of control. We compare the proposed nonparametric Bayesian procedure with two other commonly used MTPs. One is the error-control procedure described by Benjamini and Yekutieli (2001). This is a modification of the Benjamini and Hochberg (1995) procedure and is known to provide control under certain forms of dependence. The second is the pFDR estimation approach described by Storey (2002) where the tuning parameter involved in the estimation process is optimally chosen using bootstrap cross-validation.

The simulation setup is exactly the same as in the previous section. However, we compare the procedures only when there is moderately strong correlation ($\rho = 0.5$) among the hypotheses. For each of the four scenarios defined by the two types of correlation structures and the two different values of π_0, we generate $M = 1000$ samples. For each sample, we estimate the optimal nominal level γ for thresholding the p-values. For the nonparametric Bayesian procedure this is done by inverting the predicted FDP curve, $\widehat{\text{FDP}}_{\text{NB}}(\gamma)$ say, at a target control level α, that is, by solving the equation $\widehat{\text{FDP}}_{\text{NB}}(\gamma) = \alpha$. For the ith sample, we denote this threshold by $\hat{\gamma}_{\text{NB},\alpha,i}$. For Storey's procedure, the nominal threshold level is found by inverting the estimated pFDR curve, say $\widehat{\text{pFDR}}_{\text{ST}}(\gamma)$, at the target level α, that is, by solving $\widehat{\text{pFDR}}_{\text{ST}}(\gamma) = \alpha$. We denote this threshold by $\hat{\gamma}_{\text{ST},\alpha,i}$. The corresponding Benjamini–Yekutieli threshold, $\hat{\gamma}_{\text{BY},\alpha,i}$, is obtained following the multi-step testing procedure described in the article by Benjamini and Yekutieli (2001).

For each sample, we compute the FDP, $\text{FDP}_i(\hat{\gamma}_{\alpha,i})$ corresponding to a threshold $\hat{\gamma}_{\alpha,i}$. Then an estimate of the FDR at level α is given by

$$\widehat{\text{FDR}}(\alpha) = M^{-1} \sum_{i=1}^{M} \text{FDP}_i(\hat{\gamma}_{\alpha,i}).$$

The thresholds for the three methods and the associated FDR estimates are denoted by $\hat{\gamma}_{\text{NB},\alpha,i}$, $\hat{\gamma}_{\text{ST},\alpha,i}$, $\hat{\gamma}_{\text{BY},\alpha,i}$ and $\widehat{\text{FDR}}_{\text{NB}}(\alpha)$, $\widehat{\text{FDR}}_{\text{ST}}(\alpha)$, $\widehat{\text{FDR}}_{\text{BY}}(\alpha)$, respectively. Table 2 gives the obtained FDR values for the three procedures. The values under the column "FDR" are the target FDR levels α. All procedures provide desired control. However, the Benjamini–Yekutieli procedure is quite conservative for the intraclass structure and has a performance similar to the Bayes procedure for the AR(1) structure. The Bayes procedure is very accurate for the intraclass structure. Storey's procedure is conservative in all cases, particularly for the AR(1) structure. From the table, the Bayes procedure has the optimal performance under both types of correlation structure. In the AR(1) structure, the correlation among the test statistics is assumed to be of the AR(1) form, which translates into a different form of correlation under the nonlinear p-value transformations. However, it is expected that the qualitative property, that hypotheses with indices close to each other have stronger correlation than those with indices far apart, still holds. The Bayesian procedure is able to adapt to this form through the mixture of the AR(1) correlation structure in the skew-normal mixing kernel. Table 3 gives the average nominal levels for the three procedures for different target FDP values. The levels are often too low compared to the expected true nominal level for Storey's and the B-Y procedures.

As in traditional hypothesis testing, the gain in terms of error-control under the null can result in significant improvement in power properties of an MTP. The Bayesian procedure potentially can provide large increase in power over the procedures that do not model the dependence in situations where the test statistics are indeed strongly correlated. However, the trade-off for this additional flexibility is the significant increase in computing time. For a dataset with $m = 1000$, it took about

Table 2. Obtained FDR from three different multiple testing procedures: nonparametric Bayes (NB), Storey (2002) (ST), Benjamini–Yekutieli (BY)

Correlation	π_0	FDR	FDR_{NB}	FDR_{ST}	FDR_{BY}
Intraclass	0.90	0.05	0.051	0.032	0.020
		0.10	0.103	0.069	0.057
		0.20	0.197	0.115	0.105
		0.30	0.295	0.207	0.168
	0.95	0.05	0.050	0.033	0.016
		0.10	0.105	0.076	0.068
		0.20	0.203	0.160	0.128
		0.30	0.292	0.230	0.198
AR(1)	0.90	0.05	0.045	0.017	0.044
		0.10	0.088	0.045	0.082
		0.20	0.158	0.105	0.161
		0.30	0.268	0.190	0.256
	0.95	0.05	0.043	0.019	0.044
		0.10	0.082	0.042	0.089
		0.20	0.167	0.113	0.171
		0.30	0.255	0.191	0.269

Table 3. The true nominal level γ and those obtained from the Bayesian method (γ_{NB}), Storey's method (γ_{ST}), and Benjamini–Yekutieli procedure (γ_{BY})

Correlation	π_0	FDR	γ	γ_{NB}	γ_{ST}	γ_{BY}
intraclass	0.90	0.05	0.0049	0.0051	0.0030	0.0023
		0.10	0.0192	0.0201	0.0072	0.0055
		0.20	0.0456	0.0448	0.0235	0.0211
		0.30	0.0755	0.0718	0.0472	0.0397
	0.95	0.05	0.0031	0.0031	0.0024	0.0016
		0.10	0.0075	0.0080	0.0058	0.0042
		0.20	0.0175	0.0180	0.0148	0.0100
		0.30	0.0320	0.0311	0.0222	0.0171
AR	0.90	0.05	0.0008	0.0007	0.0001	0.0007
		0.10	0.0026	0.0022	0.0007	0.0023
		0.20	0.0050	0.0036	0.0027	0.0039
		0.30	0.0080	0.0067	0.0046	0.0069
	0.95	0.05	0.0007	0.0006	0.0001	0.0006
		0.10	0.0012	0.0009	0.0006	0.0010
		0.20	0.0030	0.0025	0.0017	0.0026
		0.30	0.0050	0.0041	0.0028	0.0043

three minutes to predict the FDP using the Bayesian procedure, whereas the procedures that ignore dependence took less than a second. In many applications, the preferred correlation structure may be block-diagonal. For such structures, one can make parallel updates of the blocks and hence the computing time will be essentially the same as that of the updating of the largest block.

4.3 Analysis of Kidney Data

The data used to test our model come from an analysis of isografted kidneys from brain-dead donors. The data can be obtained from the National Center for Biotechnology Information (NCBI) database and related details about the experiment are in the article by Kusaka et al. (2005). 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 21,500 genes. We perform two-sample test between the normal rats and the brain-dead donors for the entire list of genes and the p-values and probit p-values of the two-sample tests are used for modeling. There are six rats in each group.

We fit the skew-normal mixture model for the kidney data and constrain one of the components to be N(0, 1). The AR(1) correlation structure seems to be more reasonable for these data. Of course such model assumptions should be made in the context of what is reasonable for the particular scientific problem. However, the mixture model makes the final conclusions less sensitive to the choice of the particular correlation structure. Due to the high dimension of the data, we generate a chain of length 2000 after a burn-in of 4000. At convergence, there are nine alternative components that are significant, of which only four had mixing proportion bigger than 0.001. The estimated proportion corresponding to the standard normal component was $\hat{\pi}_0 = 0.72$. The parameters of the significant mixture components along with the proportions are given in Table 4. The correlation parameter is estimated to be about 0.22. From Figure 3(a), the estimated model fits the observed data very well. The FDP curve is shown in Figure 3(b) for a range of standard nominal levels.

From the curve, for target FDP less than 1%, 5%, and 10%, the corresponding nominal levels for individual tests are approximately 0.00005, 0.0006, and 0.0023, respectively. The nominal levels are similar to those found in the simulation for the AR(1) structure along with $\pi_0 = 0.9$. Setting the FDP to 1%, there are 49 genes that were declared significant. Most of the genes (14 genes) known to have differential expression in brain-dead donors (Kusaka et al. 2005) are included in the set of 49 genes. The significant up-regulated genes at 1% FDP level were Lipocalin 2 (IDREF = 18,122), Calgranulin B (IDREF = 10,644), insulin-like growth factor binding protein 1 (IDREF = 1638), inhibin beta-B-subunit gene (IDREF = 2340), Fc receptor, IgG low affinity III (IDREF = 14,325), and FK 506 binding protein 5 (IDREF = 8499). Significant down-regulated expressions are found in Amphiphysin complete (IDREF = 9643), Slc15a2 (IDREF = 13,754), and Jagged 1 (IDREF = 1859). Some of the other genes that are observed to have somewhat differential expression for the brain-dead donor, such as Kcnn2, Birc2, and BCL2-protein 3, were not included in the list of genes at 1% FDP but were included at the 10% level. One reason for this is that the original experiment is a 2×2 factorial

Table 4. Parameters for the estimated skew-normal mixture components for kidney data

Component	μ	ω	λ	π
1	0.00	1.00	0.00	0.71
2	−0.90	1.91	2.40	0.01
3	−1.79	1.12	0.22	0.06
4	−2.53	1.01	0.77	0.10
5	−2.85	1.05	0.24	0.11

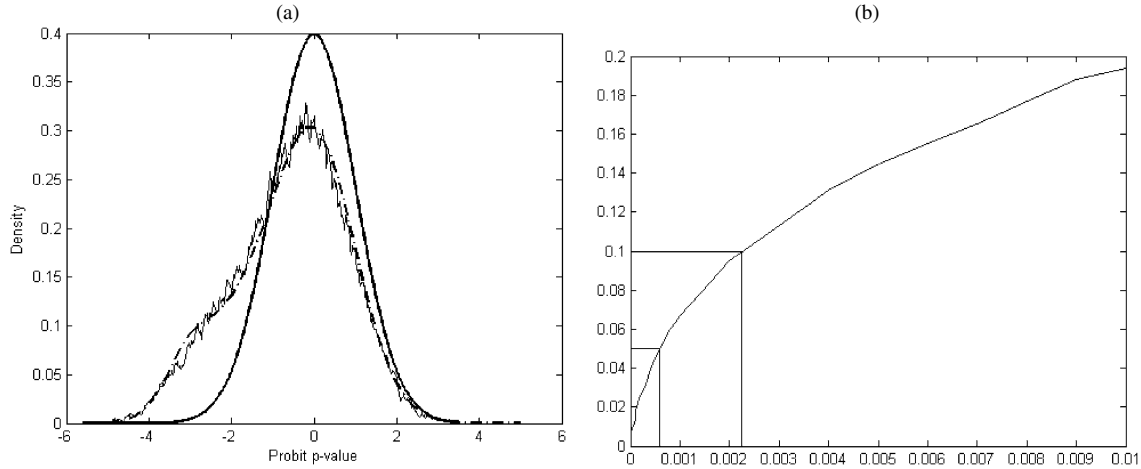

Figure 3. (a) Skew-normal mixture fit (dashed) to the Histogram polygon (jagged) of the probit p-values for the kidney data. The theoretical null distribution, the standard normal density (solid line) is also shown for reference. (b) Predicted FDP curve for the kidney data. The reference nominal levels for target FDP equal to 5% and 10% are shown.

design, and time of collection with respect to transplantation is another factor. We notice that the Kcnn2 and Birc2 have significantly smaller p-values (< 0.0001) for a two-sample test between the one-hour brain-dead donors and the rest (for Kcnn2) or between the one-hour normal donor and the rest (for Birc2). For BCL2-protein 3, rat sample number 10 is an outlier and after removing the outlier the test is significant at the nominal level corresponding to 5% FDP. Since the skew-mixture finds a substantial proportion of the alternative components, it may be prudent to use a larger FDP value for control and hence obtain a larger group of potentially important genes. The proportion of genes found significant at the 10% FDP level is about 15%.

5. CONCLUSION

We have proposed a nonparametric Bayesian procedure for multiple testing which directly models the joint distribution of the p-values via a flexible mixture model. In the simulation studies (some of which are not reported here), the procedure seems to perform remarkably well in terms of predicting the FDP curve for a given sample. This in turn results in superior control of the FDR under dependence and hence provides increase in power over the more conservative methods. Like most nonparametric Bayesian procedures, the method is computationally intensive, but the gain in statistical power will outweigh the computational cost in most applications. Like in most model-driven Bayesian procedures, sensitivity to the model assumptions and prior assumptions needs to be checked.

APPENDIX: PROOFS

Proof of Theorem 1

By Bayes's theorem, for any $\mathbf{a} = (a_1, \ldots, a_m) \in \{0, 1\}^m$, we have

$$\mathrm{E}\left[\sum_{i=1}^m I_i(1-H_i)\Big|\mathbf{I}=\mathbf{a}\right] = \sum_{i=1}^m a_i \mathrm{P}(H_i = 0|\mathbf{I}=\mathbf{a}) = \pi_0 \sum_{i=1}^m a_i \frac{\mathrm{P}(\mathbf{I}=\mathbf{a}|H_i=0)}{\mathrm{P}(\mathbf{I}=\mathbf{a})}.$$

If $\mathbf{I} = \mathbf{a}$, then clearly $\max(R, 1) = \sum_{i=1}^m a_i + \prod_{i=1}^m (1 - a_i)$. Consequently,

$$\begin{aligned}\mathrm{FDR}(\gamma) &= \pi_0 \sum_{\mathbf{a}\in\{0,1\}^m} \mathrm{P}(\mathbf{I}=\mathbf{a}) \frac{\sum_{i=1}^m a_i \mathrm{P}(\mathbf{I}=\mathbf{a}|H_i=0)/\mathrm{P}(\mathbf{I}=\mathbf{a})}{\sum_{j=1}^m a_j + \prod_{j=1}^m (1-a_j)} \\ &= \pi_0 \sum_{\mathbf{a}\in\{0,1\}^m} \frac{\sum_{i=1}^m a_i \mathrm{P}(\mathbf{I}=\mathbf{a}|H_i=0)}{\sum_{i=1}^m a_i + \prod_{i=1}^m (1-a_i)} \\ &= \pi_0 \sum_{i=1}^m \sum_{\mathbf{a}\in\mathcal{B}_i^m} \frac{\mathrm{P}(\mathbf{I}=\mathbf{a}|H_i=0)}{\sum_{i=1}^m a_i},\end{aligned}$$

proving the first part of the result.

If the observations are exchangeable, then (H_i, I_i), $i = 1, \ldots, m$, are exchangeable as well. Let, for any pair (i, j), $\mathcal{B}_{ijuv}^m$ denote the set of all m-dimensional binary vectors with u at the ith position and v at the jth position where $u, v \in \{0, 1\}$. Noting that under exchangeability

$$\sum_{\mathbf{a}\in\mathcal{B}_{ij10}^m} \frac{\mathrm{P}(\mathbf{I}=\mathbf{a}|H_i=0)}{\sum_{i=1}^m a_i} = \sum_{\mathbf{a}\in\mathcal{B}_{ij01}^m} \frac{\mathrm{P}(\mathbf{I}=\mathbf{a}|H_j=0)}{\sum_{i=1}^m a_i},$$

we have that $b_i(\gamma) = b_j(\gamma)$. This leads to the second part of the result.

Finally, note that under independence, for any $\mathbf{a} \in \mathcal{B}_1^m$, $\mathrm{P}(\mathbf{I} = \mathbf{a}|H_1) = \mathrm{P}(X_1 < \gamma|H_1 = 0)\mathrm{P}(I_2 = a_2, \ldots, I_m = a_m)$, and thus $b_1(\gamma) = \gamma \mathrm{E}[(Z+1)^{-1}]$, where $Z = R - I_1 = I_2 + \cdots + I_m \sim \mathrm{Binomial}(m-1, F(\gamma))$. It is easy to check that if $Z \sim \mathrm{Binomial}(m, p)$, then $E[(Z+1)^{-1}] = m^{-1}p^{-1}\mathrm{P}(Z^* > 0)$, where $Z^* \sim \mathrm{Binomial}(m, p)$. This gives $\mathrm{E}[(1 + R - I_1)^{-1}] = (mF(\gamma))^{-1}\mathrm{P}(R > 0)$, which implies the final assertion.

Proof of Theorem 2

If the density of the probit p-value Y is $q(y; \mu, \omega, \lambda)$, then the p-value $X = \Phi(Y)$ has density given by

$$\frac{q(\Phi^{-1}(x); \mu, \omega, \lambda)}{\phi(\Phi^{-1}(x))} = 2\sigma^{-1} \frac{e^{-z^2/2}\Phi(-\lambda z)}{e^{-(\mu+\omega z)^2/2}},$$

where $z = z(x) = \omega^{-1}(\Phi^{-1}(x) - \mu)$. Since $z(x)$ is an increasing function of x, it is enough to investigate when $r(z) = \exp\{\frac{1}{2}(\omega^2 - 1)z^2 + \mu\omega z\}\Phi(-\lambda z)$ is decreasing in z. If $\omega < 1$, then $\lim_{z\to-\infty} r(z) = 0$, and hence the density cannot be decreasing in this case.

First consider the case $\omega > 1$. If $\lambda \le 0$, then $\lim_{z\to\infty} r(z) = \infty$, and hence the density cannot be decreasing again in this case. Therefore,

let $\omega > 1$, $\lambda > 0$. Observe that $r(z)$ is decreasing, that is, $\frac{d}{dz}\log r(z) \le 0$ for all z if and only if $(\omega^2-1)z + \mu\omega - \lambda H_\Phi(\lambda z) \le 0$ for all z. The last assertion is equivalent to $\lambda^{-2}(\omega^2-1)u + \mu\omega\lambda^{-1} - H_\Phi(u) \le 0$ for all u, that is, $\mu \le \lambda\omega^{-1}\inf\{H_\Phi(u) - \lambda^{-2}(\omega^2-1)u : u \in \mathbb{R}\} = \lambda\omega^{-1}\varphi(\lambda^{-2}(\omega^2-1))$.

In view of Sampford (1953), the hazard function H_Φ of standard normal distribution is convex and the right side is finite if and only if $\lambda^{-2}(\omega^2-1) \le 1$, that is, $\lambda \ge \sqrt{\omega^2-1}$. Indeed, under this condition, the right side is nonnegative.

Finally consider the case $\omega = 1$. In this case, $r(z) = e^{\mu z}\Phi(-\lambda z)$. If $\lambda < 0$, then by the estimate $\Phi(-t) \le t^{-1}\phi(t)$, it follows that $\lim_{z\to-\infty} r(z) = 0$, which again rules out decreasing density. If $\lambda = 0$, clearly the condition for decreasing density is $\mu \le 0$, so assume $\lambda > 0$. By arguments similar to those used in the case $\omega > 1$, $r(z)$ is decreasing if and only if $\mu \le \lambda\inf\{H_\Phi(u) : u \in \mathbb{R}\} = 0$.

Proof of Proposition 1

Using (9), conditionally on $\boldsymbol{\mu}$, $\boldsymbol{\omega}$, and $\boldsymbol{\lambda}$, we have that $Y_i = \mu_i + \omega_i\delta_i|Z_0| + \omega_i\sqrt{1-\delta_i^2}Z_i$. Thus we can write $Y_i = \tau(H_i, \mu_i, \omega_i, \lambda_i, Z_0, Z_i)$, where $\tau(x_1,x_2,x_3,x_4,x_5,x_6) = x_1x_2 + x_1x_3\frac{x_4}{\sqrt{1+x_4^2}}|x_5| + \frac{x_3}{\sqrt{1+x_4^2}}x_1x_6$. Since $(H_i, \mu_i, \omega_i, \lambda_i)$ are iid and independent of $(Z_0, \mathbf{Z})$, it follows that Y_i's are strictly stationary if and only if $\mathbf{Z}$ is a strictly stationary process. Due to Gaussianity, the process $\mathbf{Z}$ is strictly stationary if and only if $\mathbf{R}$ is a stationary correlation matrix.

The following expressions of inverse and determinant of AR(1) and intraclass correlation matrices, used in the computation in Section 4, are well known in the literature.

Remark A.1. For the intraclass correlation matrix $\mathbf{R} = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}'$, the determinant is given by $\det(\mathbf{R}) = (1-\rho)^{m-1}(1+m\rho-\rho)$ and the inverse is given by $\mathbf{R}^{-1} = (1-\rho)^{-1}\mathbf{I} - \frac{\rho}{(1-\rho)(1+m\rho-\rho)}\mathbf{1}\mathbf{1}'$.

For the AR(1) correlation matrix $\mathbf{R} = ((\rho^{|i-j|}))$, the determinant is given by $\det(\mathbf{R}) = (1-\rho^2)^{m-1}$ and the inverse is given by $\mathbf{R}^{-1} = (1-\rho^2)^{-1}((r^{ij}))$, where $r^{11} = r^{mm} = 1$, $r^{ii} = 1+\rho^2$ for $i = 2,\ldots,(m-1)$, and for $i \ne j$,

$$r^{ij} = \begin{cases} -\rho, & \text{if } |i-j| = 1, \\ 0, & \text{if } |i-j| > 1. \end{cases}$$

[Received August 2010. Revised April 2011.]

REFERENCES

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178. [1210]

Azzalini, A., and Dalla Valle, A. (1996), "The Multivariate Skew-Normal Distribution," *Biometrika*, 83, 715–726. [1210,1211]

Bayarri, M. J., and Berger, J. O. (2000), "*p*-Values for Composite Null Models," *Journal of the American Statistical Association*, 95, 1127–1142. [1209]

Bazan, J. L., Branco, M. D., and Bolfarine, H. (2006), "A Skew Item Response Model," *Bayesian Analysis*, 1, 861–892. [1212]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the FDR: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57, 289–300. [1208,1214]

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188. [1208,1214]

Clarke, S., and Hall, P. (2009), "Robustness of Multiple Testing Procedures Against Dependence," *The Annals of Statistics*, 37, 332–358. [1208]

Dalla Valle, A. (2004), "The Skew-Normal Distribution," in *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, ed. M. G. Genton, Boca Raton, FL: Chapman & Hall/CRC. [1211]

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing," *Journal of the American Statistical Association*, 99, 96–104. [1208,1209]

——— (2007), "Size, Power and False Discovery Rates," *The Annals of Statistics*, 35, 1351–1377. [1208]

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160. [1208]

Escobar, M., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588. [1211]

Farcomeni, A. (2007), "Some Results on the Control of the False Discovery Rate Under Dependence," *Scandinavian Journal of Statistics*, 34, 275–297. [1208]

Finner, H., Dickhaus, T., and Roters, M. (2007), "Dependency and False Discovery Rate: Asymptotics," *The Annals of Statistics*, 35, 1432–1455. [1208]

Genovese, C., and Wasserman, L. (2002), "Operating Characteristics and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society, Ser. B*, 64, 499–517. [1208]

——— (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061. [1208]

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. [1208]

Genton, M. G. (ed.) (2004), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Boca Raton, FL: Chapman & Hall/CRC. [1210]

Ghosal, S., and Roy, A. (2011), "Identifiability of Proportion of Null Hypotheses in Mixture Models for *p*-Value Distribution in Multiple Testing," *Electronic Journal of Statistics*, 5, 329–341. [1211]

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*, eds. N. Balakrishnan et al. *IMS Collection*, Vol. 1, Beechwood, OH: Institute of Mathematical Statistics. [1210]

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537. [1208]

Hopkins, A. M., Miller, C. J., Connolly, A. J., Genovese, C., Nichol, R. C., and Wasserman, L. (2002), "A New Source Detection Algorithm Using the False-Discovery Rate," *The Astronomical Journal*, 123, 1086–1094. [1208]

Ishwaran, H., and Zarepour, M. (2000), "Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models," *Biometrika*, 87, 371–390. [1212]

Kusaka, M., Yamada, K., Kuroyanagi, Y., Terauchi, A., Kowa, H., Kurahashi, H., and Hoshinaga, K. (2005), "Gene Expression Profile in Rat Renal Isografts From Brain Dead Donors," *Transplantation Proceedings*, 37, 364–366. [1215]

MacEachern, S. N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–228. [1212]

Miller, C. J., Genovese, C., Nichol, R. C., Wasserman, L., Connolly, A., Reichart, D., Hopkins, A., Schneider, J., and Moore, A. (2001), "Controlling the False Discovery Rate in Astrophysical Data Analysis," *The Astronomical Journal*, 122, 3492–3505. [1208]

Pawitan, Y., Calza, S., and Alexander, P. (2006), "Estimation of False Discovery Proportion Under General Dependence," *Bioinformatics*, 22, 3025–3031. [1208]

Robins, J. M., van der Vaart, A., and Ventura, V. (2000), "Asymptotic Distribution of *p*-Values in Composite Null Models," *Journal of the American Statistical Association*, 95, 1143–1156. [1209]

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003), "A New Class of Multivariate Skew Distributions With Applications to Bayesian Regression Models," *Canadian Journal of Statistics*, 31, 129–150. [1211]

Sampford, M. R. (1953), "Some Inequalities on Mill's Ratio and Related Functions," *Annals of Mathematical Statistics*, 24, 130–132. [1217]

Sarkar, S. K. (2002), "Some Results on False Discovery Rate in Stepwise Multiple Testing Procedures," *The Annals of Statistics*, 30, 239–257. [1208]

——— (2004), "FDR-Controlling Stepwise Procedures and Their False Negatives Rates," *Journal of Statistical Planning and Inference*, 125, 119–137. [1208]

——— (2006), "False Discovery and False Non-Discovery Rates in Single-Step Multiple Testing Procedures," *The Annals of Statistics*, 34, 394–415. [1208]

——— (2007), "Step-Up Procedures Controlling Generalized FWER and Generalized FDR," *The Annals of Statistics*, 35, 2405–2420. [1208]

Scott, J., and Berger, J. O. (2005), "An Exploration of Aspects of Bayesian Multiple Testing," *Journal of Statistical Planning and Inference*, 136, 2144–2162. [1208]

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society, Ser. B*, 64, 479–498. [1208,1209,1214]

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-Value," *The Annals of Statistics*, 31, 2013–2035. [1208]

Storey, J. D., and Tibshirani, R. (2003), "Statistical Significance for Genomewide Studies," *Proceedings of the National Academy of Sciences USA*, 100, 9440–9445. [1208]

Tang, Y., Ghosal, S., and Roy, A. (2007), "Nonparametric Bayesian Estimation of False Discovery Rates," *Biometrics*, 63, 1126–1134. [1210]