Two-group Poisson-Dirichlet mixtures for multiple testing

Francesco Denti

Department of Statistics, University of California, Irvine

email: fdenti@uci.edu

and

Michele Guindani

Department of Statistics, University of California, Irvine

email: michele.guindani@uci.edu

and

Fabrizio Leisen

School of Mathematics, Statistics and Actuarial Sciences, University of Kent, Canterbury, UK

email: F.Leisen@kent.ac.uk

and

Antonio Lijoi

Department of Decision Sciences, Bocconi University, Milan Italy

Bocconi Institute of Data Science and Analytics (BIDSA)

email: antonio.lijoi@unibocconi.it

and

W. Duncan Wadsworth

Microsoft, Redmond, Washington

email: duwads@microsoft.com

and

Marina Vannucci

Department of Statistics, Rice University, Houston, Texas

email: marina@rice.edu

Summary: The simultaneous testing of multiple hypotheses is common to the analysis of high-dimensional data sets. The two-group model, first proposed in Efron (2004), identifies significant comparisons by allocating observations to a mixture of an empirical null and an alternative distribution. In the Bayesian nonparametrics literature, many approaches have suggested using mixtures of Dirichlet Processes in the two-group model framework. Here, we investigate employing mixtures of two-parameter Poisson Dirichlet Processes (2PPD) instead, and show how they provide a more flexible and effective tool for large-scale hypothesis testing. Our model further employs non-local prior densities to allow separation between the two mixture components. We obtain a closed form expression for the exchangeable partition probability function of the two-group model, which leads to a straightforward MCMC implementation. We compare the performance of our method for large-scale inference in a simulation study and illustrate its use on both a prostate cancer dataset and a case-control microbiome study of the gastrointestinal tracts in children from underdeveloped countries who have been recently diagnosed with moderate to severe diarrhea.

Key words: Bayesian nonparametrics, Microbiome analysis, Multiple testing, Poisson–Dirichlet process, Two-group model

1. Introduction

The availability of high-dimensional data in domains as diverse as genomics, imaging, and astronomy, has brought the necessity to screen a large number of hypotheses simultaneously. Here, we focus on the two-group modeling framework (Efron, 2004). To illustrate, we assume that the observations are suitably defined difference scores $z_i, i = 1, \ldots, n$ over a large number of distinct hypotheses. The two-group model assumes that the z_i's are drawn either from a null (f_0) or a non-null (f_1) distribution, i.e., each score is drawn from a mixture,

$$z_i \sim f = (1-\rho)f_0 + \rho f_1, \tag{1}$$

for some weight $\rho \in (0,1)$, and some probability (density) functions f_0 and f_1. The null component is typically assumed standard normal; however, the true null distribution may differ from the theoretical null, e.g., due to limited sample size or unaccounted correlation. Thus, Efron proposes the estimation of an "empirical null" distribution to adequately capture the range of parameter values coherent with the null hypothesis and accordingly evaluate each testing decision.

In Bayesian nonparametrics, the Dirichlet process (DP) has been extensively used to provide flexible estimates of f_0, or f_1, or both, as well as for clustering the z_i's into common "expression" levels (Do et al., 2005; Dahl and Newton, 2007; Kim et al., 2009; Kottas and Fellingham, 2012). **?** develop a flexible hierarchical nonparametric approach where f_0 is assigned a Normal distribution with unknown mean and variance, whereas f_1 is a location mixture of normals. One appealing feature of the two-group model is that the resulting inference is immediately amenable to interpretation in a decision theoretic framework. For example, Efron (2004) describes a local version of the false discovery rate (*local fdr*), which represents the posterior probability that a difference score z_i is generated according to the null hypothesis, $fdr(z_i) = (1-\rho)\, f_0(z_i)/f(z_i)$. The selection of interesting scores is conducted by flagging all z_i's such that $fdr(z_i) < \alpha$, $\alpha \in (0,1)$, allowing control of the

Benjamini–Hochberg FDR (Benjamini and Hochberg, 1995) at level α. More generally, the decision problem could minimize loss functions that compound expected false positive and false negative decisions. The optimal decision would then lead to thresholding the posterior probability of the alternative (e.g., see Muller et al., 2006).

In this manuscript, we investigate the use of a mixture prior of two–parameter Poisson–Dirichlet (2PPD) processes in lieu of the commonly used DPs. The 2PPD process, also known as the *Pitman-Yor* process, is a generalization of the DP and is characterized by two parameters: a "concentration" parameter θ (analogous to the single parameter of the DP), and a "discount" parameter σ. The additional parameter allows for more flexible clustering behavior than the DP and can be used to tune the reinforcement mechanism of large clusters (Lijoi et al., 2007). We show how the proper choice of σ can be used to model the empirical null distribution f_0 and the uncertainty related to the non-null distribution in the two-group model, leading to improved testing procedures. Our modeling framework further employs non-local prior densities for the base measure of the random probability measures under the alternative hypothesis to allow better separation between the two mixture components. We derive the expression of the exchangeable partition probability function (EPPF), induced by the proposed two-group 2PPD mixture process and observe that, conditional on the assignment of the observations to the null or the alternative hypothesis, the respective random partitions are independent. This property conveniently facilitates posterior inference obtained via MCMC algorithms, which take into account the conditional independence of the partitions. By means of a simulation study, we discuss the performance of our method with respect to the commonly used mixture of DPs and existing state-of-the-art approaches for large-scale multiple comparison problems. We also illustrate the use of the proposed 2PPD processes mixture model on two publicly available datasets: a well-known Prostate cancer dataset (Singh et al., 2002) and one collected from a recent microbiome study (Pop et al., 2014). In the latter case, the aim was to characterize the microbial composition of the gastrointestinal tracts of children from underdeveloped countries who have been diagnosed with moderate to severe diarrhea. Our study suggests that mixture of DPs should be used with some caution in large scale multiple-testing, and that the use of 2PPD processes could lead to improved operating characteristics.

2. A review of the 2PPD process

In this Section we provide an overview of the 2PPD process with particular regard to its use for density estimation and its clustering properties. Let $Z_1, \ldots, Z_n$ be a sample of n data measurements (e.g. raw observations or summary statistics), drawn from a sequence of exchangeable random elements $Z_1, Z_2, \ldots$, taking value in a complete and separable metric space $\mathbb{Z}$ endowed with its Borel σ-algebra $\mathscr{Z}$. By virtue of the de Finetti representation theorem,

$$\begin{aligned} Z_i \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} \qquad\qquad i = 1, \ldots, n, \\ \tilde{p} &\sim Q, \end{aligned} \tag{2}$$

for any $n \geqslant 1$, and for $\tilde{p}$, a random probability measure, with distribution Q defined on the space $\mathscr{P}(\mathbb{Z})$ of probability measures on $\mathbb{Z}$. In a Bayesian framework, Q represents the prior distribution and the model is said to be parametric whenever Q degenerates on a finite dimensional subspace of $\mathscr{P}(\mathbb{Z})$; otherwise, the model is denoted as nonparametric.

Here, we consider the 2PPD process for the random probability measure $\tilde{p}$, which can be represented almost surely as an infinite mixture, i.e., $\tilde{p} = \sum_{k=1}^{\infty} \tilde{w}_k \, \delta_{Y_k}$, where δ_c denotes the point mass at c, the $\tilde{w}_k$'s are random weights obtained as $\tilde{w}_1 = V_1$ and $\tilde{w}_k = V_k \prod_{j=1}^{k-1} (1 - V_j)$, $k \geqslant 2$ with $V_j \overset{ind}{\sim} \text{Beta}(1-\sigma, \theta + j\sigma)$, $j \geqslant 1$ (*stick-breaking construction; Pitman, 1995)*, for some $\sigma \in [0,1)$ and $\theta > -\sigma$. The Y_k's are random locations in $\mathbb{Z}$, independent of the weights $\tilde{w}_k$'s, and assumed as random draws from a non-atomic *base measure* P^*, i.e., $Y_k \overset{\text{iid}}{\sim} P^*$, $k \geqslant 1$, which represents the prior expected value of the random distribution $\tilde{p}$, i.e., $\mathbb{E}[\tilde{p}(A)] = P^*(A)$ for any $A \in \mathscr{X}$. We should note that the 2PPD process is also well defined for $\sigma < 0$ and $\theta = r|\sigma|$, with r being an integer; however, in such case the process reduces to the parametric Fisher model (Ghosal and van der Vaart, 2017). Hereafter, we will use $Z_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$, with $\tilde{p} \overset{d}{=} \text{2PPD}(\sigma, \theta, P^*)$, $i = 1, \dots, n$ to indicate a sample from a 2PPD with parameters σ and θ, and base measure P^*. If $Z_1, \dots, Z_n$ is a realization from an exchangeable sequence driven by a 2PPD process, there is a positive probability of ties, i.e., $\mathbb{P}[Z_i = Z_j] > 0$ for any $i \neq j$. This *clustering* property often motivates the use of the 2PPD process in statistical applications, e.g. to model data from heterogeneous populations.

The clustering behavior of the 2PPD process can also be investigated by considering the exchangeable partition probability function (EPPF), which characterizes the probability that $Z_1, \dots, Z_n$ are partitioned into K distinct clusters with respective sizes $n_1, \dots, n_K$. For the 2PPD process, such probability is $\Pi_K^{(n)}(n_1, \dots, n_K) = \frac{\prod_{j=1}^{K-1}(\theta + j\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{K}(1-\sigma)_{n_j - 1}$ for any choice of positive integers $n_1, \dots, n_K$ such that $\sum_{i=1}^{K} n_i = n$, with $K \in \{1, \dots, n\}$ and $(a)_q = \Gamma(a+q)/\Gamma(a)$, for any non-negative integer q. The expression highlights how the values of the parameters σ and θ affect the clustering structure induced by the 2PPD process. It is well-known that if K_n denotes the number of distinct values recorded in a sample $Z_1, \dots, Z_n$ of an exchangeable sequence drawn according to a $\text{2PPD}(\sigma, \theta)$ process, then $K_n/n^\sigma \to S_{\sigma,\theta}$ as $n \to \infty$ (almost surely) for some positive random variable $S_{\sigma,\theta}$ when $\sigma \in (0, 1)$(see Theorem 3.8 in Pitman, 2002). When $\sigma = 0$, we recover the clustering behavior of the Dirichlet process, obtaining $K_n/\log n \to \theta$ as $n \to \infty$ (almost surely). Hence, the larger σ is, the larger the number of clusters. Moreover, σ controls the reinforcement of the partition, i.e., the ability of big clusters to attract even more observations, as highlighted by the predictive distribution of the 2PPD process,

$$\mathbb{P}[Z_{n+1} \in A \mid Z_1, \ldots, Z_n] = \frac{\theta + \sigma\, K_n}{\theta + n - 1}\, P^* + \sum_{j=1}^{K_n} \frac{n_j - \sigma}{\theta + n - 1}\, \delta_{Z_j^*}(A),$$

where the probability that a new observation is assigned to an existing cluster, and assumes value Z_j^*, $j = 1, \ldots, K_n$, is proportional to $n_j - \sigma$. Therefore, values of σ close to 1 favor the formation of a large number of clusters, most of which are singletons (Lijoi et al., 2007).

Finally, we consider the variability of realizations from a 2PPD process around the base measure P^*. The variance of the process is $\mathrm{Var}[\tilde{p}(A)] = \frac{1-\sigma}{\theta+1}\, P^*(A)[1 - P^*(A)]$, for any $A \in \mathscr{X}$ and $j = 0, 1$. Large values of σ correspond to random probability measures which are more concentrated around the base measure P^*. Therefore, one should expect that the empirical distribution function of any sample $Z_1, \ldots, Z_n$ drawn from a 2PPD process with high values of σ, $F_n(b) = \tilde{p}(\infty, b] = \sum_{k=1}^{\infty} \tilde{w}_k\, \delta_{Z_k^*}(\infty, b]$, would be characterized by a large number of weights $\tilde{w}_k$ of similar size. In the next Sections we will exploit these properties to guide the use of the 2PPD process in the two-group model for multiple testing.

3. Methods

3.1 *A two-group 2PPD model*

The different clustering behavior that the 2PPD process exhibits as a function of σ can be exploited for distinguishing between the null and alternative distributions in the two-group model. More precisely, we first rewrite model (2) as the two–component mixture,

$$\tilde{p} = (1 - \rho)\, \tilde{p}_0 + \rho\, \tilde{p}_1, \tag{3}$$

where $\tilde{p}_j \sim 2\mathrm{PPD}(\sigma_j, \theta_j, P_j^*)$ represents the unknown distribution under the null and the alternative hypotheses, for $j = 0$ and $j = 1$, respectively. Similarly as in (1), the mixture weight ρ is a random variable independent of the $\tilde{p}_j$'s and takes values in $[0, 1]$. We further introduce an auxiliary binary random variable γ_i, $i = 1, \ldots, n$, such that $Z_i \sim \tilde{p}_0$ if $\gamma_i = 0$

and $Z_i \sim \tilde{p}_1$ if $\gamma_i = 1$. Thus, conditionally on the $\gamma_i^{\prime} s$, we can rewrite (2)–(3) as

$$\begin{aligned} Z_i \mid \gamma_i &\overset{\text{ind}}{\sim} \tilde{p}_{\gamma_i}, \qquad i = 1, \dots, n, \\ \gamma_i \mid \rho &\overset{\text{iid}}{\sim} \text{Bernoulli}(\rho), \\ \tilde{p}_{\gamma_i} &\sim 2\text{PPD}(\sigma_{\gamma_i}, \theta_{\gamma_i}, P^*_{\gamma_i}), \end{aligned} \tag{4}$$

with $\tilde{p}_0$ and $\tilde{p}_1$ independent, and assuming a Beta(a, b) distribution on ρ. The hyperparameters a and b influence the proportion of discoveries and can be tuned according to the problem at hand. In genomic studies, one may want to enforce sparsity of discoveries, with prior expected proportions $\mathbb{E}[\rho] = \frac{a}{a+b}$ between 1% and 10% of the total number of hypotheses. A lower value of $\mathbb{E}[\rho]$ typically results in lower posterior probabilities of the alternative, although the relative ranking of the posterior probabilities is overall preserved.

We exploit the properties of the 2PPD process discussed in Section 2 and propose to specify the hyperparameters of the null and non-null random probability measures in (4) as follows. In accordance with Efron's idea that the empirical null distribution should capture only small departures from the theoretical null, we let $\tilde{p}_0$ concentrate around the theoretical null. Furthermore, we assume that there's no good model *a priori* for the non-null distribution. Therefore, $\tilde{p}_1$ is allowed to vary more freely on the space of the alternative distributions. Under the null distribution, the process should encourage the creation of a large number of clusters each composed by few observations, so that the empirical distribution well approximates the theoretical null. For the non-null distribution, we should expect a more uneven distribution of the realizations. Based on those considerations, we propose to set $\sigma_0 > \sigma_1$. We will discuss how such a choice might help discriminating between the null and the alternative distribution in the multi-comparison problem.

We conclude this Section by considering the joint partition structure induced by model (4) for a sample $Z_1, \dots, Z_n \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$. Let $\Pi^{(n)}_{K,j}(n_1, \dots, n_K)$ denote the EPPF of process

$\tilde{p}_j$, $j = 0, 1$, that is the probability that n observations are assigned to K different clusters of sizes $(n_1, \ldots, n_K)$. For notational simplicity, we assume that $\Pi^{(n)}_{K+1,j}(n_1, ..., n_K, 0) \equiv \Pi^{(n)}_{K,j}(n_1, \ldots, n_K)$, for any $j = 0, 1$ and $n_1, \ldots, n_K \geqslant 1$ such that $\sum_{i=1}^K n_i = n$. Then the following result provides the EPPF of the mixture of 2PPD processes as below:

Proposition 1. *The EPPF associated to the mixture of 2PPD processes in* (4) *is given by:*

$$\Pi^{(n)}_K(n_1, \ldots, n_K) = \frac{1}{(a+b)_n} \sum_{\boldsymbol{i} \in \times_{j=1}^K \{0, n_j\}} (a)_{|\boldsymbol{i}|} (b)_{n-|\boldsymbol{i}|} \times \\ \Pi^{(|\boldsymbol{i}|)}_{K_0,0}(i_1, \ldots, i_K)\, \Pi^{(n-|\boldsymbol{i}|)}_{K_1,1}(n_1 - i_1, \ldots, n_K - i_K) \quad (5)$$

where $\boldsymbol{i} = (i_1, \ldots, i_K)$, $|\boldsymbol{i}| = i_1 + \cdots + i_K$, $K_0 = card\{j : i_j = n_j\}$ *and* $K_1 = K - K_0$. *If* $i_k = n_k$ *or* $i_k = 0\, \forall k$, *we assume* $\Pi^{(n)}_K(i_1, \ldots, i_K) = 1$.

See Web Appendix B for a proof. Direct use of (5) is far from trivial. Nonetheless, the expression lends itself to an interesting interpretation: conditional on the assignment of the clusters to either $\tilde{p}_0$ or $\tilde{p}_1$, the respective random partitions are still independent. This remark is useful for devising a suitable computational algorithm for posterior inference.

3.2 *Bayesian hierarchical two-group mixture model*

In many applications, the discreteness of the realizations of the 2PPD process may be considered inadequate. Thus, in lieu of (4), it is often common to assume for a sample $Z_1, \ldots, Z_n$ a hierarchical mixture model with continuous components, i.e.

$$Z_i | \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, \quad \text{with} \quad \tilde{p} = (1-\rho) \int k_0(Z_i, \vartheta)\, \tilde{p}_0(d\vartheta) + \rho \int k_1(Z_i, \vartheta)\, \tilde{p}_1(d\vartheta), \qquad (6)$$

that is the two-group model is characterized by a null and non-null distributions which are each defined as a 2PPD process mixture. Here, $f_{\tilde{p}}(Z_i)$ is the random density induced by the random probability measure $\tilde{p}$, while $k_j : \mathbb{Z} \times \Theta \to \mathbb{R}^+$, $j = 0, 1$ are general kernels such that for $\boldsymbol{\vartheta} \in \Theta$ and some σ–finite measure λ on $(\mathbb{Z}, \mathscr{Z})$ one has $\int_{\mathbb{Z}} k_j(x, \boldsymbol{\vartheta})\, \lambda(\mathrm{d}x) = 1$, $j = 0, 1$. For our purposes, it is convenient to set $\mathbb{Z} = \mathbb{R}$ and let λ coincide with the Lebesgue measure on $\mathbb{R}$ so that the previous model defines a prior on the space of density functions on $\mathbb{R}$. By conditioning on the auxiliary group indicator variables γ_i, $i = 1, \ldots, n$, we can rewrite model (6) as a hierarchical Bayes *two-group 2PPD process mixture*,

$$\begin{aligned} Z_i \mid \boldsymbol{\vartheta_i}, \gamma_i &\overset{\text{ind}}{\sim} k_{\gamma_i}(Z_i \mid \boldsymbol{\vartheta}_i), \qquad i = 1, \ldots, n \\ \boldsymbol{\vartheta}_i \mid \gamma_i, \tilde{p} &\overset{\text{ind}}{\sim} \tilde{p}_{\gamma_i}, \\ \gamma_i \mid \rho &\overset{\text{iid}}{\sim} \text{Bernoulli}(\rho), \\ \rho &\sim \text{Beta}(a, b) \\ \tilde{p}_{\gamma_i} &\sim \text{2PPD}(\sigma_{\gamma_i}, \theta_{\gamma_i}, P^*_{\gamma_i}), \end{aligned} \tag{7}$$

where $\boldsymbol{\vartheta}_i$ may indicate either a scalar or a vector parameter. In general, $k_0(\cdot)$ and $k_1(\cdot)$ could be different. Here, we assume $k_{\gamma_0}(\cdot) = k_{\gamma_1}(\cdot) = k(\cdot)$ to be a Normal kernel and set $\boldsymbol{\vartheta_i} = (\mu_i, \tau_i^2)$. For notational simplicity, in (7) we have omitted additional hyperparameters which may feature in the kernel function $k(\cdot)$ but are not relevant for the decision problem and thus are assigned separate priors.

We conclude the specification of the two-group model (7) by discussing the choice of the base measures P_0^* and P_1^*. On the one hand, we achieve flexible estimation of the so-called "empirical null" distribution by setting $P_0^*\left(\mu, \tau^2\right) = \pi\left(\mu\right) \times \pi\left(\tau^2\right) = \mathcal{N}\left(0, 1\right) \times IG\left(a_0, b_0\right).$ where the parameters of the IG on τ^2 are chosen so to allow relatively small deviations from the theoretical null distribution. For example, by assuming $a_0 = 5$, $b_0 = 0.2$, the induced marginal distribution on Z_i has only slightly fatter tails than the standard normal.

Moreover, P_0^* and P_1^* should not have significantly overlapping supports, i.e. they should assign high probability to regions of the parameter space that are consistent with the null and the alternative hypotheses, respectively. In the Bayesian multiple hypotheses testing framework, this requirement has sometimes been advocated to ensure enough *separation* between the null and the alternative models. Thus, we first model P_1^* as a symmetric bimodal mixture of Normal-Inverse Gamma (NIG) distributions, as $P_1^* = \frac{1}{2}\text{NIG}(-|m_1|, k_1, \alpha_1, \beta_1) +$

$\frac{1}{2}\mathrm{NIG}(|m_1|, k_1, \alpha_1, \beta_1)$, with $m_1 \in \mathbb{R}$, and $k_1, \alpha_1, \beta_1 \in \mathbb{R}^+$. Marginally,

$$\pi\left(\mu_1|m_1\right) = \frac{1}{2}\left[\sqrt{\frac{\beta_1}{\alpha_1}} t_{2\alpha_1} - |m_1|\right] + \frac{1}{2}\left[\sqrt{\frac{\beta_1}{\alpha_1}} t_{2\alpha_1} + |m_1|\right].$$

We further achieve separation in the multiple hypotheses testing problem by modeling the location parameter m_1 with a non-local prior (NLP), i.e. a prior that assigns vanishing density to small neighborhoods of the null hypothesis (Johnson and Rossell, 2010). Several types of NLP have been proposed in the literature. See, for instance, Rossell and Telesca (2017). Here, we adopt an r-th moment (MOM) prior for m_1, with

$$\pi_{MOM}\left(m_1; 0, \kappa^2, r\right) = \frac{m_1^{2r}}{\xi} \frac{e^{-m_1^2/2\kappa^2}}{\sqrt{2\pi\kappa^2}}, \tag{8}$$

where ξ is the normalizing constant, and we write $m_1 \sim NLP_{MOM}\left(0, \kappa^2, r\right)$. Specific hyperparameter specifications will be detailed in Section 4. Here, we only note that the non-local prior specification in P_1^* should provide enough separation from the origin to ensure good estimation of the posterior probability of the alternative. Finally, the other parameters of the 2PPD processes are set such that $\theta_0 = \theta_1$ and $\sigma_0 > \sigma_1$. In general, θ_0 and θ_1 are chosen relatively small, in order to enforce coarser clustering structures, especially under the alternative hypothesis. Typically, in Dirichlet-Process two-groups models, $\theta_0 = \theta_1 = 1$ (see, e.g. Do et al., 2005). From the discussion at the end of Section 2, it follows that realizations of the 2PPD null process are expected to be more concentrated around the base measure. In the next Sections we will investigate the effect of different choices for the parameter values of the 2PPD processes for the multiple comparison problem.

3.3 *Posterior inference*

Posterior inference for model (4) or (7) relies on Markov Chain Monte Carlo techniques since the posterior distributions are not available in closed form. Our primary interest is in the group indicators γ_i's, which uniquely identify the random probability measure from which the data Z_i's were generated, and, correspondingly, the probability of group membership, ρ.

For the sampling of the γ_i's, we exploit the independence of the random partitions implied by the EPPF (5) of the proposed mixture of 2PPD processes. More specifically, if $Z_1, \ldots, Z_n$ are a random sample from (4) and P_j^*, $j = 0, 1$ are non-atomic base measures with common support, then $\mathbb{P}[Z_i = Z_j \mid \gamma_i \neq \gamma_j] = 0$ for $i \neq j$. Thus, all the Z_i's in a cluster are generated by the same 2PPD process. The details of the MCMC algorithms are provided in the Web Appendix A. In particular, we employ a split-merge move to speed up computations for large sample sizes (Dahl, 2005). The computational burden of the MCMC algorithm increases for higher values of either θ_0, θ_1, σ_0 or σ_1 due to the increased number of latent clusters generated by the 2PPD process. A discussion of the computational efficiency of a plain Pólya-Urn sampler versus the split-merge implementation is also provided in the Web Appendix D.

Posterior inference on the weight ρ in (4) is conducted by means of post-MCMC analysis, by approximating the posterior expected value $\mathbb{E}[\rho \mid \text{data}]$ using auxiliary indicators, say $\boldsymbol{\gamma}_t^* = (\gamma_{1,t}^*, \ldots, \gamma_{K^{(t)},t}^*)$, which denote if cluster $k \in 1, \ldots K^{(t)}$ at iteration $t = 1, \ldots, T$ is a realization from $\tilde{p}_0$ or $\tilde{p}_1$. More precisely, if we denote by $B < T$ the burn-in period of the chain, we can compute the following Monte Carlo approximation of the posterior expected value $\mathbb{E}[\rho \mid \text{data}] \approx \frac{1}{T-B} \sum_{t=B+1}^{T} \frac{a + \sum_{k=1}^{K^{(t)}} n_{k,t}(1-\gamma_{k,t}^*)}{a+b+n}$.

Similarly, the posterior probability that an observation belongs to the non-null group can be obtained from the MCMC output as $PP_i^1 = p(\gamma_i = 1 \mid \text{data}) \approx \frac{1}{T-B} \sum_{t=B+1}^{T} \gamma_{i,t}$, where the $\gamma_{i,t}$'s indicate the MCMC draws of the component indicators γ_i's. Then, a score Z_i is considered significant if the corresponding PP_i^1 is larger than a threshold, say κ, chosen to control the Bayesian FDR at a pre-assigned $\alpha \times 100\%$ level , $BFDR\left(\kappa\right) = \frac{\sum_{\nu=1}^{V} \left(1 - \mathrm{PP}_{\mathrm{i}}^{1}\right) I\left(\mathrm{PP}_{\mathrm{i}}^{1} > \kappa\right)}{\sum_{\nu=1}^{V} I\left(\mathrm{PP}_{\mathrm{i}}^{1} > \kappa\right)} < \alpha$ (Newton et al., 2004; Muller et al., 2006).

4. Applications

4.1 *Simulation study*

We investigate the performance of the Bayesian hierarchical 2PPD mixture modeling framework described in (6)–(7) for large-scale multiple hypothesis testing by means of a simulation study under $S = 5$ scenarios. More specifically, we simulate z-scores from mixture (1), where $f_0(z) = \mathcal{N}(z \mid 0, \sigma_s^2)$. We set $\sigma_s^2 = 1$ for $s = 1, \dots 4$. For the fifth scenario, we set $\sigma_5^2 = 1.5$ to model the effect of hidden correlation among observations and of the association with unobserved covariates, that may lead to departures from standard Gaussianity. For f_1 we choose:

- **Scenario 1**: $f_1(z) = 0.67 \cdot \mathcal{N}(z| - 3, 2) + 0.33 \cdot \mathcal{N}(z|3, 2)$,
- **Scenario 2**: $f_1(z) = \mathcal{N}(z \mid u, 1)$ with $u \sim \text{Uniform}(2, 4)$,
- **Scenario 3**: $f_1(z) = \mathcal{N}(z \mid u, 1)$ with $u \sim \text{Uniform}([-4, -2] \cup [2, 4])$,
- **Scenario 4**: $f_1(z) = \text{Gamma}((-1)^v \cdot z \mid a, b)$ with $a = 4$, $b = 1$ and $v \sim \text{Bernoulli}(0.5)$,
- **Scenario 5**: $f_1(z) = 0.5 \cdot \mathcal{N}\left(z|5, 1\right) + 0.5 \cdot \mathcal{N}\left(z| - 5, 1\right)$,

i.e. f_1 is assumed asymmetric unimodal (scenario 1), symmetric bimodal (scenarios 2), asymmetric bimodal (scenario 3) and symmetric bimodal with fat tails (scenario 4 and scenario 5), thus mimicking typical high-dimensional testing situations. An illustrative plot of data generated under the five scenarios is provided in the Web Appendix C. In all scenarios, we set $\rho = 0.05$, since typically only a small proportion of the comparisons is expected to be significant in large-scale inference hypothesis testing. Each simulation includes $n = 1,000$ simulated scores and is replicated 30 times to allow quantification of posterior uncertainty and of the frequentist operating characteristics of the testing procedures.

[Table 1 about here.]

For model fitting, we employ the mixture model (6)–(7), where we assume $k(\cdot \mid \theta_i) =$

Normal$(\cdot \mid \boldsymbol{\vartheta}_i)$, with $\boldsymbol{\vartheta}_i = (\mu_i, \tau_i^2)$. The base measure of the 2PPD process $\tilde{p}_0$ is chosen as described in Section 3.2, with $a_0 = 5$, $b_0 = 0.2$. For P_1^*, we set $k_1 = 1/3$, $\alpha_1 = 1$, $\beta_1 = 1$. A NLP_{MOM} prior is assumed for m_1, with $r = 3$ and $\kappa = 2$. For the parameters characterizing the clustering behavior of the 2PPD process priors, we investigate the effect of different choices of (σ_0, σ_1) on the inference, with $\sigma_0 > \sigma_1$. More specifically, here we report the inference for the following values for the pair (σ_0, σ_1): $(0.75, 0)$, which corresponds to assuming a DP on the non-null component; in addition to $(0.75, 0.1)$, $(0.75, 0.25)$, $(0.9, 0.25)$ to investigate the effect of decreased prior uncertainty, $Var(\tilde{p})$, on the components of the two-group 2PPD mixture. We further set the concentration parameters $\theta_0 = \theta_1 = 1$ (Do et al., 2005). For the Beta prior on ρ, we set $a = 1$ and $b = 9$. For each dataset, the MCMC algorithm was run for 2,500 iterations after a 2,500 iterations burn-in period. The evaluation of posterior convergence was conducted using standard Bayesian convergence diagnostics on the chains of the traceable parameters, m_1 and ρ, by monitoring the number of group components and by inspecting the estimated densities of the null and non-null processes.

We compare the performance of our modeling approach with five alternative methods for large-scale hypothesis testing: (a) a two-group DP mixture model, which can be seen as a special case of the modeling framework proposed here, obtained by setting $\sigma_0 = \sigma_1 = 0$, with a non-local prior on the base measure for the alternative distribution (b) the local false discovery rate of Efron (2004); (c) the Benjamini and Hochberg procedure (BH, Benjamini and Hochberg, 1995); (d) the empirical Bayes mixture model of Muralidharan (2012), which allows simultaneous estimation of the effect size and of the local false discovery rate, and (e) the empirical Bayes semi-parametric approach of **?**.

For each simulation replicate, results were compared using several performance measures: the Matthews Correlation Coefficient (MCC), which can be computed from a confusion matrix as MCC $= (TP \times TN - FP \times FN)/\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$, where TP, TN, FP, and FN are the number of true positive, true negative, false positive and false negative results, respectively; the F1 score, $2\mathrm{TP}/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$; as well as precision, specificity, accuracy and the area under the curve (AUC) of the corresponding receiver operating characteristic curve. For each simulation, we identify significant scores by controlling the Bayesian false discovery rate (Newton et al., 2004), the local false discovery rate (Efron, 2004) and the frequentist false discovery rate (Benjamini and Hochberg, 1995) at the 10% level.

[Table 2 about here.]

In Table 1 we report the performance metrics achieved in the different simulation scenarios as a function of the combinations of hyperparameters of the 2PPD process. Overall, the performance of the proposed 2PPD process is similar, as long as $\sigma_1 < \sigma_0$. Higher values of σ_0 lead to draw samples from f_0 which are closer to the theoretical null, but the implied tighter control of the variance of the null process may lead to a slightly decreased performance in some scenarios. If $\sigma_1 > \sigma_0$, the performance can deteriorate considerably.

Table 2 reports the results from the comparison with alternative multiple testing methods. Compared to our method, the method of **?** performs quite well in all scenarios except the fat-tailed one, Scenario 4, where our 2PPD model outperforms four out of five competitors. The BH procedure also performs quite well, although with slightly lower precision, in the first four scenarios. However, small departures from the standard Gaussian null assumption (scenario 5) considerably affect the performance of the BH procedure. The performance of two-group DP mixtures is impacted by the flexible modeling of both the null and alternative distribution, which leads to a relatively high number of false assignments. This result is remarkable as various types of mixture of DP processes have been often proposed for hypothesis testing in the two-group modeling framework. The results also appear fairly robust to different sample sizes (see Web Appendix E).

4.2 *Case study: Microbiome data*

We illustrate the applicability of the proposed two-group 2PPD process model on a publicly available dataset of microbial abundances from a case-controlled study on post-diarrheal disruption in children from low-income countries. The purpose of the study was to identify potential microbiota which may show positive associations with moderate-to-severe diarrhea (MSD) in the case group. Negative associations are also of interest since they may suggest potential target treatments for recovery from dysbiosis.

Stool samples were obtained from 992 children between the ages of 0 and 59 months, 508 of whom had recently suffered from moderate to severe diarrhea, with the remaining 484 children acting as age-matched controls. The samples were obtained in Mali (M), the Gambia (G), Kenya (K), and Bangladesh (B) and case/control proportions were approximately equal for each country.

[Figure 1 about here.]

Due to the nature of the sampling mechanism, the distribution of the microbiome counts is highly skewed, i.e., a few are highly abundant, whereas most microbes have low frequencies (Chen and Li, 2016). Here, we are interested in evaluating the ability of our model to identify microbiota which may be differently abundant in healthy and MSD subjects. Therefore, we employ a Negative-Binomial regression model on the taxonomic abundances y_{ij}, where $j = 1, \ldots, J_i$ indexes the microbiotic taxa, and $i = 1, \ldots, n$ indexes the samples. As it is typical when dealing with sequencing data (see, e.g., Witten, 2011), we let s_i denote an estimate of a sample-specific size factor, to take into account the different sequencing depths of the samples. Also, we let x_{ij}^{case}, x_{ij}^{age} and $x_{ij}^{country}$ denote the three available covariates for the MSD status, age and country. More specifically, $x_{ij}^{case} = 1$ for cases and $x_{ij}^{case} = 0$ for the matched controls. We adopt Gambia as the reference value for the other countries, and let x_{ij}^K, x_{ij}^B, and x_{ij}^M be dummy variables for the other countries. Then, we assume:

$$y_{ij} \overset{\text{ind}}{\sim} NB(\mu_{ij}, \alpha_j), \qquad j = 1, \ldots, J_i;\ i = 1, \ldots, n,$$

$$\log(\mu_{ij}) = \log(s_i) + \beta_{0,j} + \beta_{1,j}\, x_{ij}^{case} + \beta_{2,j}\, x_{ij}^{age} + \beta_{3,j}\, x_{ij}^{M} + \beta_{4,j}\, x_{ij}^{B} + \beta_{5,j}\, x_{ij}^{K} + \epsilon_{ij},$$

where α_j represents a taxon-specific dispersion parameter, and $\beta_{0,j}$ represents a taxon-specific effect, which captures the abundance of taxon j in the control group, and the $\beta_{k,j}$'s represent the effects of each covariate on the taxon abundance. The Negative Binomial distribution was chosen due to its flexibility over the Poisson alternative. The model was fitted using the `glmmTMB` package. To illustrate our multiple testing procedure, we consider the fixed case-control effect captured by the estimates of the coefficients $\beta_{1,j}$'s, which provide the z-scores for testing the differences in abundance between healthy and MSD subjects. A histogram of the 535 z-scores from the data is given in Figure 1. Since the estimated coefficients are a function of the original data, the independence assumption may not be satisfied if the original taxonomic abundances are correlated. Indeed, the presence of hidden correlation among the observables and unknown associations with unobserved covariates are major motivations for the two-group model formulation in Efron (2004).

In the two-group model (6)–(7), we fix the hyperparameters for the prior processes as $\theta_0 = \theta_1 = 1$, $\sigma_0 = 0.75$, $\sigma_1 = 0.10$. The specific choice for σ_0 allows small departures of the empirical null from the theoretical $\mathcal{N}(0, 1)$ distribution, while maintaining computational feasibility in the generation of the latent clusters from the null. A $\text{Beta}(1, 99)$ is chosen for ρ to further encourage sparsity of discoveries. The hyperparameters of the base measures were set as in Section 4.1. For the results provided here, we run 20,000 iterations after 20,000 iterations as burn-in. Figure 1 overlays the Monte Carlo estimates of the posterior probability of each taxon belonging to the non-null distribution to the histogram of the z-

scores. By thresholding the Monte Carlo estimate of posterior probability of the non-null process at a value corresponding to a Bayesian false discovery rate (Newton et al., 2004) of 1%, we identify a total of 74 non-null taxa. On the contrary, the BH procedure leads to 143 significant microbes, when controlling the FDR at the 1% level. The *locfdr* model detects as relevant only 6 taxa. Tables 1 and 2 in the Web Appendix F report the taxa with the highest discovery probabilities, separately for positive and negative z-scores. A close inspection of our results reveals some interesting biological findings (see Web Appendix G).

4.3 *Case study: Prostate Cancer Dataset*

To assess how our model performs in large-sample cases, we apply our methodology to the widely known *Prostate* dataset of Singh et al. (2002). See also Efron (2009). We exploit the split-merge move in the MCMC to improve computational efficiency (see Web Appendix D). The dataset is composed of 6,033 genes for 102 observations from 52 prostate cancer patients and 50 healthy men. We adopt the same prior specification as in the microbiome case study, with the exception that here we set $b = 9$, as in the simulation studies. This choice is in accordance with the discussion in Efron (2008), who suggests a proportion *a priori* of no more than 10% non-null genes for these data. Figure 2 reports the posterior probabilities of discovery for this dataset. When thresholding the BFDR at the 20% level, our method flags only 18 genes as relevant. Similarly, the *locfdr* procedure flags 19 genes. On the contrary, the BH procedure identifies 60 genes as significant, even when thresholding the FDR at the 10% level.

[Figure 2 about here.]

5. Discussion and Conclusion

We have considered the two-group model by Efron (2004) for multiple hypotheses testing and we have proposed the use of a mixture prior of two–parameter Poisson–Dirichlet processes as a flexible class of prior processes in that framework. In particular, an appropriate choice of the hyperparameters of the 2PPD processes allows the characterization of small departures from the theoretical null in the estimation of the empirical null distribution, while leaving flexibility in the modeling of the non-null distribution. We have also employed a mixture of non-local prior densities as base measure for the alternative distribution, to improve separation and facilitate the estimation and identifiability of the mixture components. The proposed approach has been shown to provide a robust testing procedure, which compares favorably with recently proposed methods for estimating the components of the two-group model, including the widely-used DP mixture models. A limitation of the procedure is related to the computing effort, since Markov chain Monte Carlo algorithms for Bayesian nonparametric models typically require considerable computational time for posterior inference. To provide an illustration, in the analysis of the Prostate cancer dataset of Section 4.3, it took approximately 56 hours to run 20,000 MCMC iterations on a Xeon(R) E5-2640 v4, 2.40GHz Linux sever, with the computational bottleneck being represented by the iterations requiring a full Pólya-Urn sampling. Variational Bayes techniques have been developed for many Bayesian nonparametric models, including the 2PPD process (see, e.g. Jordan and Blei, 2006). However, the speed up of MCMC algorithms for Bayesian nonparametric models in high-dimensional settings is still a topic of ongoing research (see, e.g., Canale et al., 2019).

A careful choice of the hyperparameters of the two-group 2PPD model is essential to ensure good operating characteristics of the testing procedures. We have followed prevailing practices and set $\theta_0 = \theta_1 = 1$ in both the simulations and the data analyses. Priors on θ_0 and θ_1 would need to incorporate constraints to facilitate the identification of the two-group components.

Finally, in our data analyses, we have proposed a two-group model for the analysis of data observed under two conditions. However, often the interest is in studying longitudinal changes of repeated measurements within a subject. Therefore, models that take into account the temporal dependence of the hypotheses are required.

Data Availability Statement

The data that support the findings in this paper are openly available in the R packages `msd16s` at `http://doi.org/10.1186/gb-2014-15-6-r76` (Pop et al., 2014) and `sda`, at `https://CRAN.R-project.org/package=sda` (Ahdesmaki et al., 2015).

References

Ahdesmaki, M., Zuber, V., Gibb, S., and Strimmer, K. (2015). *sda: Shrinkage Discriminant Analysis and CAT Score Variable Selection.* R package version 1.3.7.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **57,** 289–300.

Canale, A., Corradin, R., and Nipoti, B. (2019). Importance conditional sampling for Bayesian nonparametric mixtures.

Chen, E. Z. and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **32,** 2611–2617.

Dahl, D. B. (2005). Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* .

Dahl, D. B. and Newton, M. A. (2007). Multiple hypothesis testing by clustering treatment effects. *Journal of the American Statistical Association* **102,** 517–526.

Do, K. A., Müller, P., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **54,** 627–644.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99,** 96–104.

Efron, B. (2008). Microarrays, empirical bayes and the two-groups model. *Statist. Sci.* **23,** 1–22.

Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association* **104,** 1015–1028.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference.* Cambridge University Press.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **72,** 143–170.

Jordan, M. I. and Blei, D. M. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis* **1,** 121–143.

Kim, S., Dahl, D. B., and Vannucci, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Analysis* **4,** 707–732.

Kottas, A. and Fellingham, G. W. (2012). Bayesian semiparametric modeling and inference with mixtures of symmetric distributions. *Statistics and Computing* **22,** 93–106.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **69,** 715–740.

Muller, P., Parmigiani, G., and Rice, K. (2006). FDR and Bayesian multiple comparisons rules. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., West, M., and Smith, A. F. M., editors, *Bayesian Statistics 8*, number 1995, pages 349–370. Oxford University Press.

Muralidharan, O. (2012). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Annals of Applied Statistics* **6,** 422–438.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155–176.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102,** 145–158.

Pitman, J. (2002). *Combinatorial Stochastic Processes.* Springer-Verlag.

Pop, M., Paulson, J. N., Bravo, H. C., et al. (2014). Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biol* **15,** R76.

Rossell, D. and Telesca, D. (2017). Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association* **112,** 254–265.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1,** 203–209.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *Annals of Applied Statistics* **5,** 2493–2518.

Supporting Information

The Web Appendices referenced in Sections 3, 4, and 5 are available with this paper at the Biometrics website on Wiley Online Library. The code is openly available online at `https://github.com/mguindanigroup/twogroup2PPD`.

Received May 2019. *Revised* —- —-. *Accepted* —- —-.

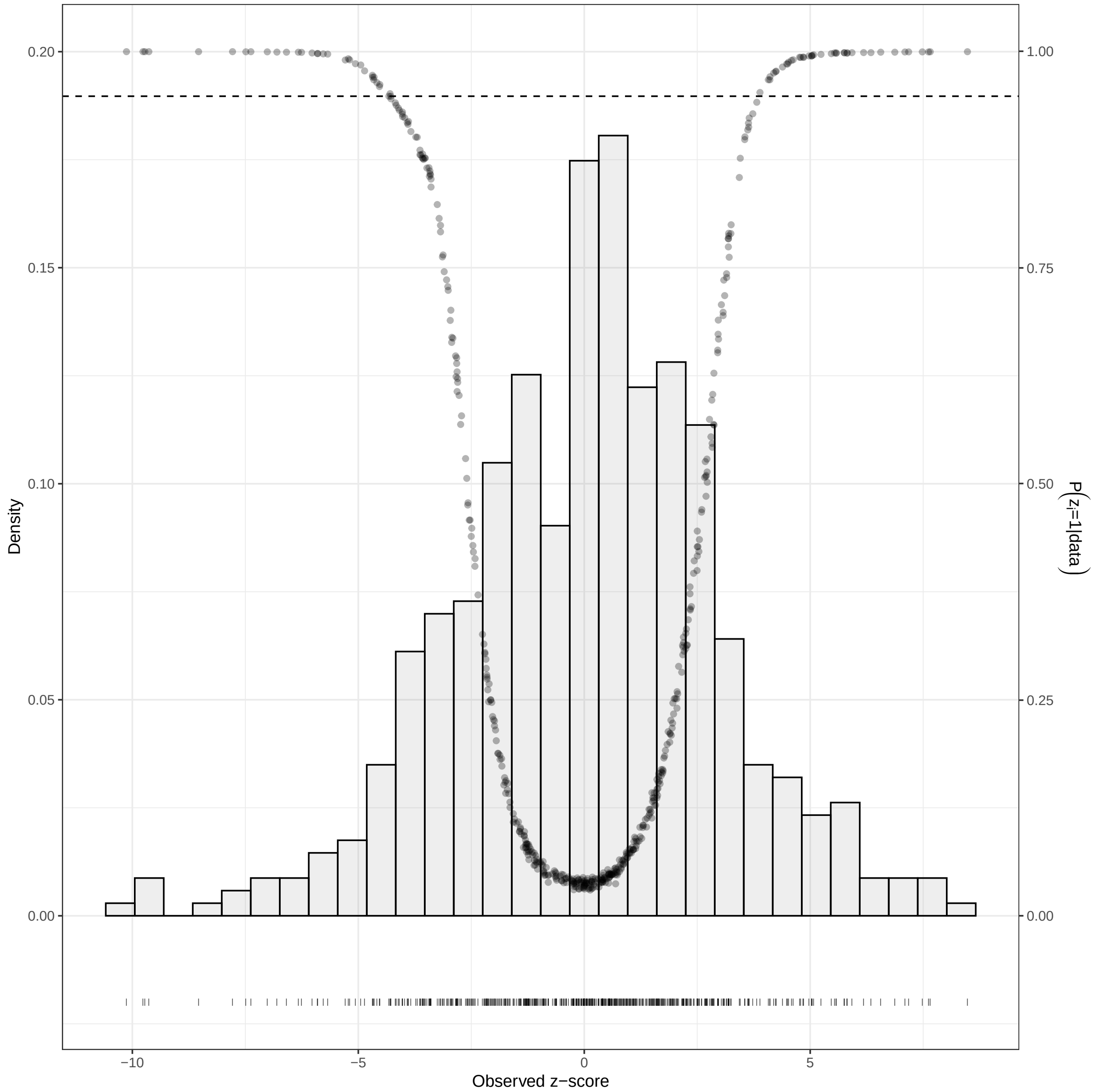

Figure 1. Microbiome data case study: Histogram of 535 z-scores obtained from the case term (β_1) in the Negative Binomial generalized linear mixed effects model. We superimpose the posterior probabilities of the events $\{\gamma_i = 1|\boldsymbol{z}\}$ and the threshold corresponding to a Bayesian FDR of 1%.

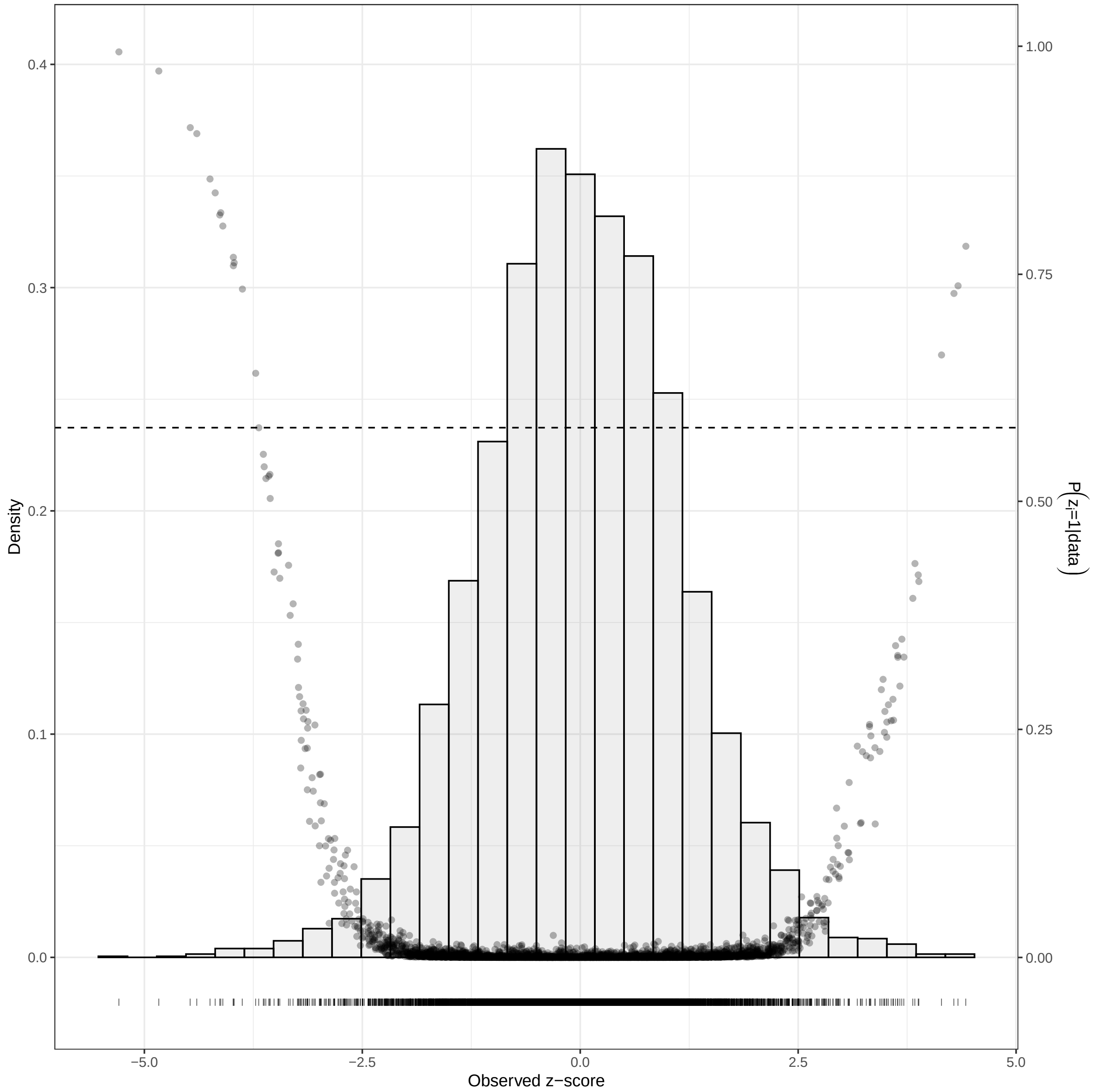

Figure 2. Prostate dataset: Histogram of 6033 z-scores obtained from a two-groups comparison. We superimpose the posterior probabilities of the events $\{\gamma_i = 1|\boldsymbol{z}\}$ and the threshold corresponding to a Bayesian FDR of 20%.

Table 1

Simulation study: sensitivity results across different settings for σ_0 and σ_1 for the five simulation scenarios considered in Section 4.1 ($\rho = 0.05$). The values in the table represent the average MCC and F_1 scores, the average precision (PRE), specificity (SPEC), accuracy (ACC) and the area under the curve (AUC) of the corresponding receiver operating characteristic curve, over 30 replicates with corresponding standard deviations between brackets.

	$\sigma_0 = 0.75$			$\sigma_0 = 0.9$
	$\sigma_1 = 0$	$\sigma_1 = 0.1$	$\sigma_1 = 0.25$	$\sigma_1 = 0.25$
Scenario 1				
MCC	0.5777 (0.0903)	0.5833 (0.0940)	0.6020 (0.0835)	0.5893 (0.0876)
F1	0.5197 (0.1125)	0.5269 (0.1169)	0.5520 (0.1051)	0.5342 (0.1095)
AUC	0.9095 (0.0245)	0.9143 (0.0224)	0.9201 (0.0253)	0.9200 (0.0207)
PRE	0.9775 (0.0357)	0.9773 (0.0333)	0.9713 (0.0346)	0.9776 (0.0328)
SPEC	0.9995 (0.0007)	0.9995 (0.0007)	0.9994 (0.0008)	0.9995 (0.0007)
ACC	0.9676 (0.0049)	0.9680 (0.0052)	0.9690 (0.0048)	0.9683 (0.0049)
Scenario 2				
MCC	0.6249 (0.0673)	0.6242 (0.0695)	0.6212 (0.0681)	0.6135 (0.0644)
F1	0.5809 (0.0831)	0.5796 (0.0855)	0.5771 (0.0835)	0.5665 (0.0808)
AUC	0.9526 (0.0218)	0.9563 (0.0185)	0.9581 (0.0170)	0.9523 (0.0183)
PRE	0.9710 (0.0338)	0.9725 (0.0373)	0.9680 (0.0396)	0.9712 (0.0384)
SPEC	0.9993 (0.0008)	0.9994 (0.0009)	0.9993 (0.0009)	0.9993 (0.0009)
ACC	0.9703 (0.0043)	0.9703 (0.0044)	0.9701 (0.0043)	0.9696 (0.0040)
Scenario 3				
MCC	0.5081 (0.0842)	0.5080 (0.0847)	0.5340 (0.0797)	0.5224 (0.0808)
F1	0.4320 (0.1053)	0.4320 (0.1056)	0.4659 (0.1001)	0.4489 (0.1018)
AUC	0.9335 (0.0235)	0.9401 (0.0238)	0.9477 (0.0180)	0.9452 (0.0209)
PRE	0.9721 (0.0438)	0.9714 (0.0425)	0.9682 (0.0402)	0.9772 (0.0360)
SPEC	0.9995 (0.0007)	0.9995 (0.0007)	0.9994 (0.0007)	0.9996 (0.0006)
ACC	0.9624 (0.0044)	0.9624 (0.0045)	0.9638 (0.0045)	0.9631 (0.0043)
Scenario 4				
MCC	0.7513 (0.0462)	0.7554 (0.0462)	0.7625 (0.0461)	0.7572 (0.0478)
F1	0.7354 (0.0538)	0.7413 (0.0528)	0.7535 (0.0518)	0.7449 (0.0542)
AUC	0.9552 (0.0162)	0.9627 (0.0119)	0.9685 (0.0107)	0.9661 (0.0087)
PRE	0.9787 (0.0264)	0.9736 (0.0284)	0.9532 (0.0312)	0.9657 (0.0289)
SPEC	0.9993 (0.0009)	0.9991 (0.0010)	0.9984 (0.0013)	0.9988 (0.0010)
ACC	0.9789 (0.0034)	0.9792 (0.0035)	0.9797 (0.0035)	0.9794 (0.0036)
Scenario 5				
MCC	0.8920 (0.0229)	0.8832 (0.0249)	0.8529 (0.0232)	0.8694 (0.0241)
F1	0.8951 (0.0219)	0.8860 (0.0241)	0.8534 (0.0235)	0.8710 (0.0238)
AUC	0.9985 (0.0010)	0.9985 (0.0010)	0.9985 (0.0011)	0.9985 (0.0011)
PRE	0.8346 (0.0300)	0.8170 (0.0334)	0.7560 (0.0330)	0.7856 (0.0326)
SPEC	0.9898 (0.0021)	0.9885 (0.0025)	0.9832 (0.0029)	0.9859 (0.0026)
ACC	0.9886 (0.0020)	0.9875 (0.0028)	0.9831 (0.0030)	0.9855 (0.0029)

Table 2

Simulation study: performance metrics for five other multiple comparison methods in the five simulation scenarios considered in Section 4.1 ($\rho = 0.05$). The values in the table represent the average MCC and F_1 scores, the average precision (PRE), specificity (SPEC), accuracy (ACC) and the area under the curve (AUC) of the corresponding receiver operating characteristic curve, over 30 replicates with corresponding standard deviations between brackets.

	DPmix		*local fdr*		Benjamini and Hochberg (1995)		Muralidharan (2012)		?	
Scenario 1										
MCC	0.2329	(0.0209)	0.5708	(0.0721)	0.6629	(0.0648)	0.5379	(0.0804)	0.5835	(0.0757)
F1	0.1980	(0.0145)	0.5067	(0.0932)	0.6427	(0.0736)	0.4643	(0.1023)	0.5251	(0.0962)
AUC	0.9053	(0.0301)	0.8869	(0.0365)	0.9237	(0.0205)	0.9242	(0.0230)	0.9230	(0.0216)
PRE	0.1113	(0.0092)	0.9897	(0.0216)	0.9141	(0.0603)	0.9915	(0.0227)	0.9825	(0.0284)
SPEC	0.6150	(0.0432)	0.9998	(0.0004)	0.9974	(0.0019)	0.9999	(0.0004)	0.9996	(0.0006)
ACC	0.6297	(0.0396)	0.9671	(0.0042)	0.9726	(0.0044)	0.9653	(0.0043)	0.9679	(0.0044)
Scenario 2										
MCC	0.2506	(0.0180)	0.6088	(0.0773)	0.6674	(0.0638)	0.5805	(0.0667)	0.6435	(0.0686)
F1	0.2037	(0.0138)	0.5578	(0.0991)	0.6486	(0.0721)	0.5194	(0.0857)	0.6048	(0.0846)
AUC	0.9524	(0.0218)	0.9231	(0.0428)	0.9544	(0.0169)	0.9668	(0.0174)	0.9762	(0.0087)
PRE	0.1140	(0.0087)	0.9796	(0.0304)	0.9129	(0.0556)	0.9895	(0.0216)	0.9698	(0.0332)
SPEC	0.6033	(0.0364)	0.9995	(0.0008)	0.9974	(0.0018)	0.9998	(0.0004)	0.9993	(0.0008)
ACC	0.6213	(0.0339)	0.9694	(0.0048)	0.9729	(0.0042)	0.9676	(0.0004)	0.9715	(0.0044)
Scenario 3										
MCC	0.2337	(0.0195)	0.5397	(0.0854)	0.6544	(0.0578)	0.4840	(0.0883)	0.5591	(0.0740)
F1	0.1948	(0.0129)	0.4708	(0.1087)	0.6342	(0.0670)	0.3973	(0.1080)	0.4970	(0.0948)
AUC	0.9400	(0.0206)	0.9069	(0.0359)	0.9500	(0.0191)	0.9481	(0.0197)	0.9481	(0.0182)
PRE	0.1085	(0.0079)	0.9759	(0.0424)	0.9089	(0.0561)	0.9901	(0.0328)	0.9707	(0.0437)
SPEC	0.5679	(0.0349)	0.9995	(0.0008)	0.9972	(0.0020)	0.9999	(0.0005)	0.9994	(0.0010)
ACC	0.5880	(0.033)	0.9641	(0.0050)	0.9710	(0.0040)	0.9611	(0.0043)	0.9652	(0.0044)
Scenario 4										
MCC	0.2671	(0.0194)	0.7080	(0.0474)	0.7849	(0.0443)	0.6840	(0.0486)	0.6831	(0.0470)
F1	0.2161	(0.0156)	0.6801	(0.0586)	0.7853	(0.0448)	0.6492	(0.0602)	0.6485	(0.0595)
AUC	0.9612	(0.0156)	0.9406	(0.0246)	0.9709	(0.0085)	0.9627	(0.0159)	0.9658	(0.0139)
PRE	0.1217	(0.0099)	0.9919	(0.0172)	0.9136	(0.0502)	0.9972	(0.0106)	0.9953	(0.0123)
SPEC	0.6288	(0.0345)	0.9998	(0.0005)	0.9965	(0.0023)	0.9999	(0.0003)	0.9999	(0.0004)
ACC	0.6459	(0.0325)	0.9758	(0.0033)	0.9812	(0.0036)	0.9741	(0.0034)	0.9741	(0.0032)
Scenario 5										
MCC	0.2671	(0.0194)	0.8632	(0.0492)	0.7861	(0.0360)	0.8506	(0.0486)	0.8879	(0.0332)
F1	0.5303	(0.0420)	0.8611	(0.0529)	0.7792	(0.0395)	0.8475	(0.0414)	0.8888	(0.0338)
AUC	0.9980	(0.0012)	0.9971	(0.0041)	0.9986	(0.0010)	0.9985	(0.0012)	0.9985	(0.0011)
PRE	0.3622	(0.0163)	0.9811	(0.0286)	0.6433	(0.0531)	0.9866	(0.0229)	0.9745	(0.0323)
SPEC	0.9058	(0.0155)	0.9992	(0.0013)	0.9705	(0.0067)	0.9994	(0.0009)	0.9988	(0.0015)
ACC	0.9104	(0.0385)	0.9878	(0.0041)	0.9716	(0.0064)	0.9867	(0.0032)	0.9899	(0.0028)