Received: 8 March 2022 | Revised: 2 October 2022 | Accepted: 24 February 2023

DOI: 10.1002/sim.9705

RESEARCH ARTICLE

Statistics in Medicine WILEY

Multiple hypothesis screening using mixtures of non-local distributions with applications to genomic studies

Francesco Denti[1] | **Stefano Peluso**[2] | **Michele Guindani**[3] | **Antonietta Mira**[4,5]

[1]Department of Statistics, Università Cattolica del Sacro Cuore, Milan, Italy

[2]Department of Statistics and Quantitative Methods, University of Milan - Bicocca, Milan, Italy

[3]Department of Biostatistics, University of California Los Angeles, California, Los Angeles, USA

[4]Faculty of Economics, Università della Svizzera italiana, Lugano, Switzerland

[5]Department of Science and High Technology, University of Insubria, Como, Italy

Correspondence
Francesco Denti, Department of Statistics, Università Cattolica del Sacro Cuore, Milan, Italy.
Email: francesco.denti@unicatt.it

The analysis of large-scale datasets, especially in biomedical contexts, frequently involves a principled screening of multiple hypotheses. The celebrated two-group model jointly models the distribution of the test statistics with mixtures of two competing densities, the null and the alternative distributions. We investigate the use of weighted densities and, in particular, non-local densities as *working* alternative distributions, to enforce separation from the null and thus refine the screening procedure. We show how these weighted alternatives improve various operating characteristics, such as the Bayesian false discovery rate, of the resulting tests for a fixed mixture proportion with respect to a local, unweighted likelihood approach. Parametric and nonparametric model specifications are proposed, along with efficient samplers for posterior inference. By means of a simulation study, we exhibit how our model compares with both well-established and state-of-the-art alternatives in terms of various operating characteristics. Finally, to illustrate the versatility of our method, we conduct three differential expression analyses with publicly-available datasets from genomic studies of heterogeneous nature.

KEYWORDS

Dirichlet process mixture, multiple hypothesis testing, non-local distributions, two-group model, weight function, weighted density

1 | INTRODUCTION

Multiple hypothesis tests are often needed in the statistical analysis of large biomedical datasets to screen whether N appropriately defined test statistics $\boldsymbol{z} = \{z_i\}_{i=1}^N$ are realizations from a given *null* model or not. For instance, screening procedures are pivotal for detecting differentially regulated genes associated with disease occurrences.[1] In this context, mixture models represent a flexible statistical tool, widely employed to cast the hypothesis testing problem in terms of selection and estimation of competing models. In this direction, Efron[2] proposed a two-group model to select and estimate an empirical null distribution and the corresponding alternative. Mixture models have also been proposed for distributions of P-values.[3] In a Bayesian framework, Do et al[4] employed Dirichlet process mixture models of Gaussian densities to describe null and alternative components. Martin and Todkar[5] developed a likelihood-based analysis of the two-group model, with a semiparametric specification of the non-null density. Muralidharan[6] proposed an empirical Bayes hierarchical mixture model to simultaneously estimate the effect size and the local or tail-area false discovery rate for each test statistic.

Arguably, the objective is to identify relevant cases generated from the alternative model, but the amount of separation between competing mixture components can crucially affect the performance of the tests. To our knowledge, the available approaches pose no control on the possible detrimental overlap between the null and the non-null distribution. Here, we propose a likelihood-based analysis of the two-group model, where the non-null distribution is explicitly chosen to improve the discriminating power of the testing procedure. More specifically, we first define a class of weighted densities obtained by rescaling a density function via an appropriately defined weight function.[7] The class includes many known distributions as special cases, like the skew Normal[8,9] and the non-local densities proposed by Johnson and Rossell.[10] Then, we suggest using non-local likelihood functions as *working* alternative distributions to enforce improved separation from the null model. The term *working* highlights that these distributions are not chosen to represent the actual distributions of the data under the alternative hypothesis, but only to improve the screening of the hypotheses. Thus, this modification of the two-group model, which results from the incorporation of available prior knowledge about the support of the data at the level of the likelihood, gives us direct control over the amount of separation between the two distributions.

The article is structured as follows. First, we introduce the concept of weighted densities and develop an easily interpretable parametric Bayesian two-group model in Section 2. A Bayesian nonparametric extension is also proposed. In Section 3, we prove how the use of a non-local likelihood leads to increased power and area under the curve (AUC), and lower Bayesian false discovery and Bayesian false omission rates with respect to a non-weighted likelihood approach. We employ a computationally efficient collapsed Gibbs sampler for estimating both the parametric and nonparametric specifications of the model. To conduct posterior inference, in Section 4 we discuss the adopted post-processing of the results and provide an estimate of the local false discovery rate (*lfdr*[2,11]), which is —additionally— constrained in $[0, 1)$, a natural requirement nevertheless often violated in the literature. We compare our methodology against established alternatives on simulated scenarios in Section 5 and on benchmark gene and proteomic expression datasets in Section 6. Section 7 discusses some potential extensions and conclusions.

2 | NON-LOCAL LIKELIHOOD TWO-GROUP MODEL

2.1 | Weighted densities and non-local distributions

Let X be a random variable with support $\mathcal{S}_X$ and probability density function $\pi(x;\eta)$. Let $w(x;\xi)$ be a non-negative function with parameters ξ, such that $\mathbb{E}_\pi[w(X;\xi)] < \infty$. Then, a (proper) *weighted density* function is defined by rescaling $\pi(x;\eta)$ via the weight function over $\mathcal{S}_X$, that is,

$$\pi_w(x;\xi,\eta) = \frac{w(x;\xi)}{\mathbb{E}_{\pi(x;\eta)}[w(X;\xi)]}\pi(x;\eta). \tag{1}$$

Weighted densities of the form (1) have been previously introduced by Rao,[7] who provides a formalization as an adjustment to enhance density specification when knowledge about the data generating mechanism is available. In the context of robust Bayesian analysis, they have been discussed in Bayarri and Berger[12] and, more recently, in Ruggeri et al.[13]

Many well-known distributions can be expressed as weighted densities characterized by specific weight functions. Trivially, a truncation of the random variable X on $[a, b] \in \mathcal{S}_X$ can be obtained by setting $w(x;\xi) = \mathbb{I}_{\{x\in[a,b]\}}$ with $\xi = (a, b)$. More complex truncations are obtainable by considering the sum of indicator functions on disjoint sets. A more elaborated example is the skew normal distribution, which is defined by weighting a Gaussian density via a Gaussian cumulative distribution function (c.d.f). One can also show that multivariate repulsive distributions belong to this family. For example, define $A = \{(s,j) : s = 1, \dots, k; j < s\}$ and let $g : \mathbb{R}^+ \to [0, M]$ be a strictly monotone differentiable function, with $g(0) = 0$, $g(x) > 0$ for all $x > 0$ and $M < \infty$. Then, with $w(x) = \min_{\{(s,j)\in A\}} g\left(||x_s - x_j||_2\right)$ we obtain the repulsive distribution of Petralia et al.[14]

This article considers another type of weighted densities: non-local densities. Non-local priors have been introduced by Johnson and Rossell:[10] these priors balance the convergence rates of the Bayes factor under the null and alternative hypotheses as the number of samples increases. Here, we recast their use as working alternative densities in a likelihood-based approach to multiple testing. A density $\pi_{NL}(x)$ is a *non-local density* on $\mathcal{S}_X^0 \subset \mathcal{S}_X$ if, for every $\varepsilon > 0$, there is a $\zeta > 0$ such that $\pi_{NL}(x) < \varepsilon$ for all $x \in \mathcal{S}_X$ for which $\inf_{x_0\in\mathcal{S}_X^0} |x - x_0| < \zeta$.

Hence, non-local densities assign a negligible amount of probability to the subspace S_X^0. Following the Bayesian literature, we will refer to a density that does not satisfy the previous definition as a *local density*.

A non-local density can be operatively defined by rescaling a local one. For example, if we consider a univariate normal distribution, the weighted density obtained by assuming $w(x; x_0, k) = (x - x_0)^{2k}$ defines the so-called moment (MOM) distribution, whereas $w(x; x_0, \xi) = \exp\left\{\sqrt{2} - \tau\nu/(x - x_0)^2\right\}$ with $\xi = (\nu, \tau)$ defines the exponential-moment density (eMOM).[15] More generally, a non-local distribution around x_0 is obtained by imposing that $w(x; \xi) \to 0$ as $x \to x_0$, regardless of the form of $\pi(x; \eta)$.[16] We exploit this behavior by employing a non-local density to identify significant observations beyond a region of irrelevance.

2.2 | Non-local likelihood and two-group model

We focus on multiple tests of N hypotheses. Let z_i denote a standardized test statistic, $i = 1, \dots, N$, and let $H_0^{(i)} : z_i \sim f_0$ be the i-th null hypothesis. This setting is typical, for example, of large-scale screening in genomics. Here, the objective is to quickly identify a few targets of interest, for example, genes that are differentially expressed across conditions. Alternative hypotheses do not typically represent a well-determined belief about the true distribution of the statistics. Still, their purpose is to help reach a conclusion about the evidence against the null. Thus, any specific distributional assumption for the alternative hypothesis, say $H_1^{(i)} : z_i \sim f_1$, can be seen as a *working* alternative distribution used to detect differentially expressed genes. In other words, the choice of f_1 should be made to improve the operating characteristics of the model.

Under the assumption of exchangeable hypotheses, one could describe the hypothesis testing problem using a two-group model mixture formulation[2] by assuming

$$z_i | \rho, f_0, f_1 \overset{i.i.d.}{\sim} f(z_i) = (1 - \rho) f_0(z_i) + \rho f_1(z_i), \tag{2}$$

where $i = 1, \dots, N$ and $\rho \in (0, 1)$ denotes the mixture weight. More specifically, let $\phi(z; \mu, \sigma^2)$ denote a normal density with mean μ and variance σ^2. A natural choice for f_0 would be the *theoretical null* $\phi(z; 0, 1)$. However, a standard Gaussian distribution may be unrealistic, especially in a multiple hypothesis testing setting. Indeed, Efron[11,17] notes that the failed model assumptions, unobserved covariates, and correlation of measurements across and within statistical units can make the null distribution effectively wider or narrower than N(0,1). Hence, following Efron's paradigm, we propose to estimate an *empirical null* distribution, which should capture departures from the theoretical null, but still be "close" to a standard Gaussian, with mean and variance estimated from the data. Therefore, we model f_0 as a normal distribution $\phi(z; \mu_0, \sigma_0^2)$, with normal-inverse gamma prior concentrated around $(0, 1)$ for $\left(\mu_0, \sigma_0^2\right)$.

In contrast, we model f_1 with a non-local distribution of the form $\pi_W(z; \xi, \eta) \propto w(z; \xi)\pi(z; \eta)$, where $w(z; \xi)$ is a weight function that induces small (zero) mass around (at) the origin, in order to enforce separation from f_0. As for the local density $\pi(z; \eta)$ we first propose a bi-modal mixture of two normals,

$$\pi\left(z; \tilde{\alpha}, \left\{\mu_j, \sigma_j^2\right\}_{j=1}^2\right) = (1 - \tilde{\alpha})\phi\left(z; \mu_1, \sigma_1^2\right) + \tilde{\alpha}\phi\left(z; \mu_2, \sigma_2^2\right),$$

with $\tilde{\alpha} \in (0, 1)$. In most cases, μ_1 and μ_2 have opposite signs, to capture the behavior of the tails. To this extent, we assume μ_1 and μ_2 to be constrained on the negative and positive semi-axis, respectively. For example, in the analysis of a genomic dataset, it may be of interest to identify under- and over-expressed groups of observations.

Let $\tilde{\theta} = \left(\rho, \tilde{\alpha}, \left\{\mu_j, \sigma_j^2\right\}_{j=0}^2, \xi\right)$ and $\tilde{\theta}_1$ be the sub-vector of parameters that pertain to the non-null distribution. Then, model (2) can be re-written as

$$z_i | \tilde{\theta} \;\sim\; (1 - \rho)\phi(z_i; \mu_0, \sigma_0^2) + \rho \frac{w(z_i; \xi)}{\tilde{\mathcal{K}}(\tilde{\theta}_1)} \left[(1 - \tilde{\alpha})\; \phi(z_i; \mu_1, \sigma_1^2) + \tilde{\alpha}\; \phi(z_i; \mu_2, \sigma_2^2)\right], \tag{3}$$

where $\tilde{\mathcal{K}}(\cdot)$ is the normalizing constant of the non-null distribution. For computational convenience, we reparameterize f_1 in model (3) as a mixture of weighted kernels:

$$f_1(z_i | \left\{\mu_j, \sigma_j^2\right\}_{j=1}^2, \alpha, \xi) = (1 - \alpha)\frac{w(z_i; \xi)\phi(z_i; \mu_1, \sigma_1^2)}{\mathcal{K}_1} + \alpha \frac{w(z_i; \xi)\phi(z_i; \mu_2, \sigma_2^2)}{\mathcal{K}_2}, \tag{4}$$

with $\mathcal{K}_j = \mathbb{E}_{\phi\left(z;\mu_j,\sigma_j^2\right)}[w(Z;\xi)]$ for $j = 1, 2$ and $\alpha = \tilde{\alpha}\mathcal{K}_2/\tilde{\mathcal{K}}$. In Section 1.1 of the Supplementary Material, we show how this equivalence holds in the general case of mixtures with J components. To provide a visual example, in the Section 3.1 of the Supplementary Material, we report a figure that illustrates how the weight function influences a priori the alternative and marginal densities, and the corresponding relevance probability. Finally, in Section 3, we will show how the induced separation between the two competing densities improves the operating characteristics of the weighted model.

2.2.1 | Model augmentation with latent membership labels

It is useful to introduce the latent allocation variables (λ_i, γ_i), $i = 1, \dots, N$, that explicitly identify the mixture components each observation is sampled from:

$$z_i|\boldsymbol{\theta} \overset{i.i.d.}{\sim} \begin{cases} \phi(z_i;\mu_0,\sigma_0^2) & \text{if} \quad \lambda_i = 0, \gamma_i = 0, \\ w(z_i;\xi)\phi(z_i;\mu_1,\sigma_1^2)/\mathcal{K}_1 & \text{if} \quad \lambda_i = 1, \gamma_i = 1, \\ w(z_i;\xi)\phi(z_i;\mu_2,\sigma_2^2)/\mathcal{K}_2 & \text{if} \quad \lambda_i = 1, \gamma_i = 2, \end{cases} \tag{5}$$

where $\boldsymbol{\theta} = \left(\left\{\mu_j, \sigma_j^2\right\}_{j=0}^2, \xi, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}\right)$, with $\boldsymbol{\Lambda} = (\lambda_1, \dots, \lambda_N)$ and $\boldsymbol{\Gamma} = (\gamma_1, \dots, \gamma_N)$. Note that γ_i is enough to identify in which of the three cases the i-th item is located, therefore λ_i has the only scope of improving model interpretability. We refer to the distribution induced by (5) as a non-local likelihood (Nollik). We complete model specification as follows:

$$\begin{aligned} z_i|\boldsymbol{\theta} &\overset{i.i.d.}{\sim} Nollik\left(\cdot|\lambda_i,\gamma_i,\left\{\mu_j,\sigma_j^2\right\}_{j=0}^2,\xi\right), \\ \gamma_i|\lambda_i,\alpha &\overset{i.i.d.}{\sim} Cat\left(1-\lambda_i, \alpha\lambda_i, (1-\alpha)\lambda_i\right), \quad \lambda_i|\rho \overset{i.i.d.}{\sim} Bern(\rho), \quad \rho \sim Beta(a_\rho, b_\rho), \\ \mu_j|\sigma_j^2 &\sim TN\left(m_j, \sigma_j^2/\kappa_j, \mathcal{M}_j\right), \quad \sigma_j^2 \sim IG(a_j, b_j), \quad j = 1, 2, \\ \left(\mu_0, \sigma_0^2\right) &\sim NIG\left(m_0, \kappa_0, a_0, b_0\right), \quad \alpha \sim Beta(a_\alpha, b_\alpha), \quad \xi \sim Q, \end{aligned} \tag{6}$$

where $\gamma_i \in \{0, 1, 2\}$, $Cat(\boldsymbol{p})$ indicates a categorical distribution with support on {0,1,2} and probability vector $\boldsymbol{p}$, *NIG* a normal-inverse gamma, and *TN* a truncated normal distribution, with $\mathcal{M}_1 = \mathbb{R}^-$, $\mathcal{M}_2 = \mathbb{R}^+$ being the truncation regions. Finally, Q is the distribution of the parameters in the weight function. Interpretability of the parameters in model (6) is straightforward. In addition, posterior simulation can be easily performed via Gibbs sampling. For further details, see Section 2.1 of the Supplementary Material.

2.3 | A Bayesian nonparametric extension

In the proposed setup, the distribution under the alternative is a *working* alternative aimed at improving the screening between relevant and irrelevant tests. From a hypothesis testing perspective, one should only require f_1 to be longer-tailed than f_0, with the non-null z_i's tending to occur far away from the origin.[11] However, the assumption of a specific parametric form under the alternative hypothesis can be too restrictive, and it may not be able to capture multi-modality or heavy-tailed behavior. Hence, to reflect the desired flexibility and lack of knowledge about f_1, we can extend (4) to a *Dirichlet Process Mixture Model* (DPMM) with non-local mixing kernels. The DPMM is defined as

$$\tilde{f}(z) = \int \varphi(z;\vartheta)G(d\vartheta), \; G \sim DP(a, H),$$

where $\varphi(z;\vartheta)$ denotes a generic kernel density parameterized by ϑ and *DP* indicates the Dirichlet process with concentration parameter a and base measure H.[18] It is well known that the realizations of a DP are almost surely discrete, $G = \sum_{j=1}^{+\infty}\omega_j\delta_{x_j}$ where $x_j \sim H$ and according to the stick-breaking representation[19] $\boldsymbol{\omega} = \{\omega_j\}_{j\geq 1} \sim SB(a)$, that is, $\omega_j = u_j\prod_{l=1}^{j-1}(1-u_l)$, $u_l \sim Beta(1, a)$ for $l \geq 1$.

Through the stick-breaking representation, we obtain a broad class of densities that favor realizations away from the origin as

$$f_1\left(z_i|\tilde{\theta}_1^{DP}\right) = \sum_{j\geq 1} \omega_j \frac{w(z_i;\xi)\phi\left(z_i;\mu_j,\sigma_j^2\right)}{\mathcal{K}_j}, \tag{7}$$

where $\tilde{\theta}_1^{DP} = \left(\{\omega_j\}_{j\geq 1}, \{\mu_j,\sigma_j^2\}_{j\geq 1}, \xi\right)$ and $\mathcal{K}_j = \mathbb{E}_{\phi\left(z;\mu_j,\sigma_j^2\right)}[w(Z;\xi)]$ for $j \geq 1$. We remark that, similarly to the parametric case, model (7) can be expressed as $f_1 = w(z,\xi)\pi(z,\eta)/\tilde{\mathcal{K}}$, that is, a non-local distortion of a nonparametric local density.

Despite the similar nomenclature, the proposed model is essentially different from the *weighted DP* of Sun et al[20] (see also References 21 and 22), where the authors employ a Dependent DP[23] in a regression framework to allow the error terms of observations with similar predictors' values to be characterized by similar distributions.

An alternative approach may assume a non-local distribution for the base measure of the prior process. However, without an appropriate choice of the concentration parameter a, such a prior choice does not prevent the resulting mixture from assigning non-negligible mass to regions around the origin.[24]

Once again, we introduce latent allocation variables that assign every observation z_i to either the null ($\lambda_i = 0,\ \gamma_i = 0$) distribution or one of the countable components of the alternative density weighted DP density ($\lambda_i = 1,\ \gamma_i = l, l \geq 1$). We collect the new parameters in $\theta^{DP} = \left(\left\{\mu_j,\sigma_j^2\right\}_{j=0}^{+\infty}, \xi, \mathbf{\Gamma}, \mathbf{\Lambda}\right)$, so we can write

$$z_i|\theta^{DP} \overset{i.i.d.}{\sim} \begin{cases} \phi(z_i;\mu_0,\sigma_0^2) & \text{if} \quad \lambda_i = 0, \gamma_i = 0, \\ \frac{w(z_i;\xi)}{\mathcal{K}_l}\phi(z_i;\mu_l,\sigma_l^2) & \text{if} \quad \lambda_i = 1, \gamma_i = l,\ \forall l \geq 1. \end{cases} \tag{8}$$

In the following, we refer to the two-group mixture (8) between the empirical null and the nonparametric alternative as a Bayesian nonparametric non-local likelihood (BNP-Nollik) model. In summary, our Bayesian nonparametric extension can be represented as

$$\begin{aligned} z_i|\theta &\overset{i.i.d.}{\sim} BNP\text{-}Nollik\left(\cdot|\lambda_i,\gamma_i,\{\mu_j,\sigma_j^2\}_{j=0}^{+\infty},\xi\right) \\ \pi\left(\gamma_i = l|\lambda_i,\boldsymbol{\omega}\right) &= \lambda_i\cdot\omega_l + (1-\lambda_i)\cdot\delta_0(l), \quad \forall l \geq 0, \qquad \lambda_i|\rho \overset{i.i.d.}{\sim} Bern(\rho), \\ \rho &\sim Beta(a_\rho, b_\rho), \quad \boldsymbol{\omega} \sim SB(a), \quad \xi \sim Q, \\ \left(\mu_0,\sigma_0^2\right) &\sim NIG\left(m_0,\kappa_0,a_0,b_0\right), \quad \left(\mu_j,\sigma_j^2\right) \sim G = NIG\left(m_G,\kappa_G,a_G,b_G\right), \end{aligned} \tag{9}$$

where we assume $\omega_0 = 0$ and m_G, κ_G, a_G, b_G denote the hyperparameters of the normal-inverse gamma distribution adopted as DP base measure for the alternative distribution. Lastly, a gamma distribution can be adopted as a prior for the concentration parameter a.

3 | PROPERTIES OF NON-LOCAL TWO-GROUP MODEL

To simplify notation, we denote with $f_1(z) = \pi(z;\eta)$ a local density for the alternative distribution and with

$$f_1^{NL}(z) = \pi_{NL}(z;\xi,\eta) = \frac{w(z;\xi)}{\mathcal{K}}\pi(z;\eta),$$

its weighted distortion as in (1), where $w(z;\xi)$ is a non-local weight function and $\mathcal{K} = \int_{-\infty}^{\infty} w(s;\xi)\ \pi(s;\eta)\ ds$ is the normalizing constant.

The screening process determines the specification of an interval $\mathcal{A} = [\underline{z},\overline{z}]$ (that is, the acceptance region) outside of which the z-scores are flagged as relevant, and the corresponding null hypotheses are rejected. Let $\mathcal{R} = \mathbb{R}/\mathcal{A}$ denote the rejection region. Without loss of generality, we assume $\overline{z} > 0$ and $\underline{z} < 0$. Following Efron,[2,11] given an acceptance region $\mathcal{A}$, we define the *Bayesian false discovery rate* as

$$FDR(\mathcal{A}) = \mathbb{P}[H_0|Z \notin \mathcal{A}] = \frac{\mathbb{P}[Z \notin \mathcal{A}|H_0](1-\rho)}{\mathbb{P}[Z \notin \mathcal{A}]} = \frac{(1-\rho)\int_{\mathcal{R}} f_0(z)dz}{\int_{\mathcal{R}}(1-\rho)f_0(z_i) + \rho f_1(z_i)dz}. \tag{10}$$

Analogously, we can also define the *Bayesian false omission rate* $FOR(\mathcal{A}) = \mathbb{P}[H_1|Z \in \mathcal{A}]$, and the *power* (sensitivity) $1-\beta(\mathcal{A}) = \mathbb{P}[Z \notin \mathcal{A}|H_1]$, where with β we indicate the type II error probability. Similar quantities can be defined when the assumed alternative distribution is non-local, that is, for $f_1^{NL}(z)$. We denote them with FDR^{NL}, FOR^{NL} and $1-\beta^{NL}$, respectively. We will show that modeling the unknown alternative with a non-local density improves these operating characteristics given a fixed mixing proportion ρ. With this in mind, we compute the differences

$$\Delta FDR(\mathcal{A}) = FDR(\mathcal{A}) - FDR^{NL}(\mathcal{A}), \qquad \Delta FOR(\mathcal{A}) = FOR(\mathcal{A}) - FOR^{NL}(\mathcal{A}), \qquad \Delta\beta(\mathcal{A}) \quad = \beta(\mathcal{A}) - \beta^{NL}(\mathcal{A}), \tag{11}$$

to provide a direct assessment of the relative performances in the unweighted and weighted versions. In Section 1.2 of the Supplementary Material, we show that all these differences simplify into the comparison of the discrepancies between the c.d.f.'s of the local and non-local distribution $\Delta F_1(z) = F_1(z) - F_1^{NL}(z)$ evaluated at the extremes of the acceptance region, implying that:

$$\Delta F_1(\overline{z}) \geq \Delta F_1(\underline{z}) \quad \Rightarrow \quad \Delta FDR \geq 0, \quad \Delta FOR \geq 0, \quad \Delta\beta \geq 0. \tag{12}$$

Thus, a sufficient condition for ensuring improved Bayesian FDR, Bayesian FOR, and power of the non-local weighted alternative is that the weighted c.d.f. is lower than its unweighted counterpart in $\overline{z}$ (so that $\Delta F_1(\overline{z}) > 0$), and higher in $\underline{z}$ (so that $\Delta F_1(\overline{z}) < 0$). This also implies that the screening procedure has a higher ROC curve and a higher AUC index (refer to Section 1.2 of the Supplementary Material for more details).

To provide a visual intuition, we display a simple example in Figure 1. Given an acceptance region $\mathcal{A} = [-2, 2]$, delimited by vertical dashed lines, we depict the local and non-local densities in red and blue, respectively. With similar colors we highlight the areas representing the power $\mathbb{P}[Z \notin \mathcal{A}|H_1]$. The non-local weight pushes the density mass away from the origin, resulting in sharper increments in the corresponding c.d.f. distant from zero.

To state a formal result, we need to postulate some reasonable regularity assumptions on the behavior of the weight function, additionally to the ones introduced in Section 2.1. We start by recalling that, generally, a weight function $w \equiv w(z;\xi)$ is non-local w.r.t. z_0 if (i) $\lim_{z\to z_0} w(z;\xi) = 0$.[16] In the hypothesis testing setting we consider here, $z_0 = 0$ represents the only interesting point where to induce vanishing mass. Thus, we require (ii) w to be weakly monotone decreasing (increasing) on the negative (positive) semi-axis. With no additional information about how to weight the support of z,

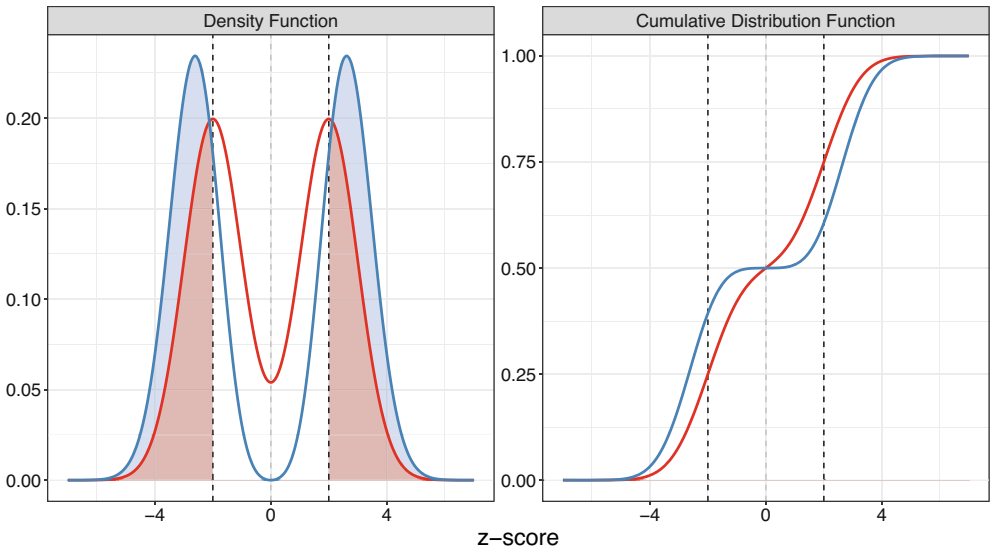

FIGURE 1 Comparison between local (red) and non-local (blue) distributions. The acceptance region $\mathcal{A} = [-2, 2]$ is highlighted by vertical dashed lines. The left panel compares the density functions, the right one the c.d.f.'s. The colored areas in the left panel represent the power $\mathbb{P}[Z \notin \mathcal{A}|H_1]$

we require (iii) $w(-z;\xi) = w(z;\xi)\,\forall z$, that is, w is an even function. The effect of the weight function has to vanish far away from the origin: an essential requirement is (iv) $w(z;\xi)\pi(z;\eta) = \mathcal{O}(\pi(z;\eta))$ as $z \to \pm\infty$, that is, the non-local density shows the same or a faster asymptotic decay than the corresponding local density. This is always the case for bounded weights. If $w(\cdot)$ satisfies the conditions (i)-(iv), we refer to it as a *proper* weight function. We can then prove the following propositions.

Proposition 1. *Consider a null hypothesis $H_0 : z \sim f_0$ characterized by an acceptance region $\mathcal{A}$. Let $w(z;\xi)$ be a proper weight function, f_1 a symmetric local density and $f_1^{NL} = \frac{w(z;\xi)}{\mathcal{K}} f_1$ its non-local distortion. Then, within the framework of the two-group model (2) assuming a fixed mixing proportion ρ, modeling the alternative distribution with f_1^{NL} rather than with f_1 ensures lower Bayesian FDR and Bayesian FOR, and higher power and AUC.*

The symmetry of the alternative distribution f_1 seems a reasonable assumption for two-tailed tests. The same result holds for one-tailed tests. If the symmetry hypothesis is removed, it is more difficult to derive a result that holds in general. However, with the introduction of a few alternative assumptions, we can prove the following:

Proposition 2. *Consider a null hypothesis $H_0 : z \sim f_0$ characterized by an acceptance region $\mathcal{A}$. Let $w(z;\xi)$ be a proper weight function, f_1 a local density and $f_1^{NL} = \frac{w(z;\xi)}{\mathcal{K}} f_1$ its non-local distortion. Define $\mathcal{S} = \{z : w(z;\xi) \leq \mathcal{K}\}$, and assume that $\mathcal{S} \subseteq \mathcal{A}$. Then, within the framework of the two-group model (2) assuming a fixed mixing proportion ρ, modeling the alternative distribution with f_1^{NL} rather than with f_1 ensures lower Bayesian FDR and Bayesian FOR, and higher power and AUC.*

We remark that these are general properties that hold every time a two-tailed test is adopted. Given an acceptance region, a two-group model with alternative non-local density and weight function satisfying (i)-(iv) has higher power, lower Bayesian FDR, and lower FOR than the corresponding local version. In Sections 1.3 and 1.4 of the Supplementary Material, we report the proofs of both propositions, concluding with an example in Section 1.5. Last, in Section 3.3 of the Supplementary Material, we discuss another advantage of the non-local specification: its robustness to prior misspecification. Specifically, the two-group model is sensitive to the choice of the distribution for ρ, which directly controls the overlap between f_0 and f_1 in the absence of other constraints. With the help of a simulation study, we show how the non-local specification helps control the number of false positives and provides more reliable estimates of the posterior probability of relevance.

4 | POSTERIOR INFERENCE

The posterior distributions $\pi(\boldsymbol{\theta}|\boldsymbol{z})$ for models (6) and (9) are not analytically tractable and we need to rely on Gibbs sampling schemes for posterior inference. For the parametric model, the full conditional distributions for ξ and $(\mu_j, \sigma_j^2)\, j = 1, 2$ require a Metropolis step. We adopt an adaptive Metropolis to improve the acceptance rate, as in Roberts and Rosenthal.[25] For the BNP Nollik model, we use the truncated representation of Ishwaran and James,[26] where the infinite sum in (7) is substituted with a sufficiently large number of mixture components J. The conditional specification allows faster computations than samplers based on Pólya Urn schemes. The samplers for both model specifications, along with comparisons in terms of the computational costs, are reported in Section 2 of the Supplementary Material.

We recover $\mathcal{A}$ by thresholding the probability of selecting the alternative distribution. In the two-group model, this is equivalent to thresholding the *lfdr*, defined as $lfdr(z) = (1-\rho)f_0(z)/f(z)$. Thus, the *acceptance region* $\mathcal{A}$ is

$$\mathcal{A} = \{z \in \mathbb{R} : lfdr(z) \geq \nu^*\} = \left\{ z \in \mathbb{R} : \frac{\rho f_1(z)}{f(z)} \leq \nu \right\}, \tag{13}$$

where $\nu = 1 - \nu^*$, $\nu \in (0, 1)$.

The fully Bayesian specification of our model allows the estimation of the parameters and functions thereof and the quantification of the uncertainty of the estimates. In particular, we are interested in the posterior probability of $H_0^{(i)}$ being rejected given by $P_1(z_i) = \mathbb{P}(\lambda_i = 1|z_i)$, that is, the probability of z_i being flagged as relevant. Once the MCMC sample is collected, we estimate $P_1(z_i)$ evaluating the ergodic mean $\hat{P}_1(z_i) = \sum_{t=1}^{T} \lambda_{it}/T$, where T is the total number of iterations and λ_{it} is the value of the chain for the parameter λ_i at the t-th MCMC step. For any $z \in \mathbb{R}$, we estimate the posterior probability of relevance $P_1(z)$ by interpolating the estimates at the observed z_i's. Alternatively, we can first estimate the densities $\hat{f}_0$ and $\hat{f}_1$ and consequently compute $\widehat{lfdr}(z)$ as defined in (13). The function $\hat{P}_1(z)$ is then obtained

as $\hat{P}_1(z) = 1 - \widehat{lfdr}(z)$. Our Bayesian model naturally constrains the range of both $lfdr(z)$ and $P_1(z)$ in $[0, 1)$, and enforces $\mathbb{P}[H_1|z = 0] = 0$, meaning that a statistic value $z = 0$ implies irrelevance almost surely. Based on the computed estimate, the hypothesis test is conducted by thresholding the function $P_1(z)$ and deriving the corresponding critical values $(\underline{z}, \overline{z})$ on the z-scores domain. We choose a threshold ν that controls, at a given level α, the Bayesian FDR (BFDR) defined in Newton et al:[27]

$$\text{BFDR}(\nu) = E(\text{FDR}|\text{Y}) = \frac{\sum_{i=1}^{N} (1 - P(z_i))\, \mathbb{I}_{\{P(z_i)>\nu\}}}{\sum_{i=1}^{N} \mathbb{I}_{\{P(z_i)>\nu\}}}. \tag{14}$$

For a specified level of α, we obtain the threshold as the minimum ν for which $\text{BFDR}(\nu) < \alpha$.

5 | SIMULATION STUDY

For the following applications, we will focus on three specific weight functions, one improper (w_0), and two proper and bounded in $[0, 1]$:

$$w_0(z; k) = z^{2k}, \qquad w_1(z; \xi, k) = 1 - e^{-\left(\frac{z}{\xi}\right)^{2k}}, \text{ and} \qquad w_2(z; \xi, k) = e^{-\left(\frac{z}{\xi}\right)^{-2k}}, \tag{15}$$

characterized by different behaviors in the way they converge to zero. For example, the weight functions w_0 and w_2 have a similar structure to the MOM and eMOM weight, respectively. However, the latter presents a sharper decay than w_1, comparable to the iMOM distribution, leading to large areas of low density for the same values of k and ξ. It is interesting to compare the two proper weight functions w_1 and w_2 in terms of their behavior around the origin. Figure 7 in the Supplementary Material shows the shape of the two weight functions for different values of $k, \xi \in \{1, 2, 3, 4\}$. We can appreciate the different effects that the two parameters induce on the chosen functions: ξ affects the functions globally, imposing a milder growth as the parameter increases. In contrast, k affects the function only in a neighborhood of the origin. Therefore, the two parameters are crucial in modeling the decay of the non-local weights and tuning the amount of separation between the null and the alternative distributions. In the following, we will set $w_0(z) = z^2$ and fix $k = 2$ in w_1 and w_2, since in our experiments the resulting power $2k$ provides a reduction of the weight in a reasonably large neighborhood of the origin sufficient to enforce the required separation. In Section 3.2 of the Supplementary Material, we report an additional simulation study that showcases the robustness of our model to several choices of the parameters ξ and k.

Once the weight functions are chosen, we can discuss the specification of the hyperprior parameters for both the parametric and nonparametric model specifications of the *working* alternative density f_1 in our model.

In the parametric case, we first assume $\xi \sim IG(a_\xi, b_\xi)$, setting $a_\xi = 20$ and $b_\xi = 57$. This choice, a priori, ensures $\mathbb{E}[\xi] = 3$, while the $\mathbb{V}[\xi] = 0.5$. As Figure 7 in the Supplementary Material shows, $\xi \approx 3$ enforces very low weight on the interval $[-1, 1]$ when combined with $k = 2$. For the mixture proportion α, we set $a_\rho = 1$ and $b_\rho = 9$, based on the assumption that only a small fraction of the observations is relevant. Moreover, we have no prior information about the proportions of the bi-modal mixture that models f_1 in Equation (3). Thus, we adopt an uniform prior imposing $a_\alpha = b_\alpha = 1$. Regarding the NIG specification for the parameters $\{\mu_j, \sigma_j^2\}_{j=1}^2$ of the alternative local distribution in (6), we set $\kappa_j = 1$, $a_j = 2$, $b_j = 5$. This implies, a priori, that $\mathbb{E}\left[\sigma_j^2\right] \approx 1.67$ and $Var\left[\sigma_j^2\right] = 6.25$. This way, we are fairly uninformative while preventing the values of the variances from assuming indefinitely large values. This choice helps prevent the estimation of extremely flat posteriors that would jeopardize the classification of the relevant observations into the under-expressed and over-expressed sets. Moreover, we adopt $m_1 = -3$ and $m_2 = 3$. For the parameters $\left(\mu_0, \sigma_0^2\right)$ of f_0 we need to specify a NIG that places most of the mass around $(0, 1)$. Therefore, we set $a_0 = b_0 = 10$ to induce a density for σ_0^2 peaked around 1. We finally set $\kappa_0 = 100$ and $m_0 = 0$, so that $\mathbb{V}\left[\mu_0|\sigma_0^2\right] = \sigma_0^2/100$.

In the nonparametric case, we truncate the stick-breaking process at $J = 30$. We then set the concentration parameter a equal to 1 and we choose a $NIG(0, 0.01, 3, 1)$ as the base measure G for the DP. These values are selected so that $\mathbb{E}\left[\mu_l|\sigma_l^2\right] = 100\sigma_l^2$ and $\mathbb{E}\left[\sigma_l^2\right] = 1/2$.[28] All the other specifications are equal to the parametric case.

We test the performance of our model on 50 datasets generated under four scenarios, adopting all the weight specifications listed in (15) under the parametric model. In addition, we also estimate the nonparametric model with the w_1

weight function. Each simulated dataset contains 1000 observations: 90% of the sample is drawn from f_0, the remaining 10% from f_1. The data generating mechanisms for the four scenarios are assumed as follows: (S1) $z_i \sim 0.90\mathcal{N}(0, 1.5) + 0.05\mathcal{N}(5, 1) + 0.05\mathcal{N}(-5, 1)$; (S2) $z_i \sim 0.90\mathcal{N}(0, 0.25) + 0.05\mathcal{N}(3, 1.5) + 0.05\mathcal{N}(-3, 1.5)$; (S3) each $z_i \sim \mathcal{N}(\gamma_i, 1)$, where γ_i is sampled from the mixture $0.90\delta_0 + 0.1\mathcal{N}(-3, 1)$. This scenario was previously proposed in Efron;[29] (S4) $z_i \sim \mathcal{N}(\gamma_i, 1)$, where γ_i is sampled from the mixture: $0.90\delta_0 + 0.10\left(0.5\mathcal{U}_{[-4,-2]} + 0.5\mathcal{U}_{[2,4]}\right)$. This scenario is similar to the one proposed by Muralidharan.[6]

We run the MCMC for 35 000 iterations, discarding the first 10 000 as burn-in period. We then thin the chains every five iterations to reduce autocorrelation, retaining a total of 5000 simulations. Visual inspection of the traceplots reveals good mixing, and the convergence of the chains was also assessed using standard MCMC diagnostics.[30] In each simulation scenario, we compute the estimate $\widehat{lfdr}(z)$. In the nonparametric case, we evaluate the posterior densities f_0 and f_1 on a grid of points at each Markov chain iteration and then consider their point-wise averages. More specifically, given T MCMC steps, we have

$$\hat{f}_0^{BNP}(z) = \frac{1}{T}\sum_{t=1}^{T}\phi(z; \mu_{0,t}, \sigma^2_{0,t}), \qquad \hat{f}_1^{BNP}(z) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J}\omega_j\phi(z; \mu_{j,t}, \sigma^2_{j,t}). \tag{16}$$

We flag the relevant hypotheses by thresholding the posterior probability of the alternative with a value that controls the BFDR (14) at a 5% level.

We compare the results obtained by our method with the `MixFDR` model,[6] the `LocFDR` model,[2] and the Benjamini-Hochberg procedure (`BH`).[31] For the first two competitors, we threshold $\widehat{lfdr}$ at 0.20, as suggested by the authors. We threshold the `BH` adjusted P-values at.05. To quantify the relative performance of the models, we compute several indices describing the operating characteristics of the procedures. More precisely, we calculate the accuracy (ACC), specificity (SPE), sensitivity (SEN), precision (PRE), and AUC of the different methods. Moreover, we compare Matthew's correlation coefficients (MCC) and the F_1 scores, defined as

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \qquad F_1 = \frac{2}{\text{SEN}^{-1} + \text{PRE}^{-1}},$$

where FP and FN denote the number of false positives and false negatives, respectively. Similarly, TP and TN denote the number of true positives and true negatives. For all indices, we report the boxplots of their distributions across the 50 Monte Carlo replications in Figures 2 and 3. Figure 2 contains indexes that summarize the overall classification performance (AUC, ACC, MCC, F_1), whereas Figure 3 showcases the different screening rates (PRE, SEN, SPE).

All the Nollik procedures lead to similar results, underlying the robustness of our proposal to different weight choices. We point out that an unbounded weight function, like w_0, is expected to over-inflate the mass far away from the origin, and therefore it is not optimal for density estimation. Nonetheless, it appears to function correctly as *working* density for estimating the posterior probability of rejection in all tests. It is also interesting to note that, despite we introduced the nonparametric specification to reflect a potential lack of knowledge regarding the true shape of the alternative distribution f_1, the parametric and nonparametric methods return similar results in all the different cases considered here. Again, this behavior suggests that, albeit the parametric specification is not optimal for density estimation, in many cases it may be sufficient to obtain competitive performances.

Overall, the simulations suggest that the Nollik models obtain very good results across all the scenarios for almost every score. Most importantly, all the scores obtained by the Nollik models are always in line, if not better, with the ones obtained by the other well-established methods. We do not observe the same robustness across scenarios in competitors. For example, `BH` presents low precision and specificity in Scenario 1, where the actual null distribution does not coincide with the theoretical one, that is, $N(0, 1)$. The `MixFDR` always presents the highest specificity, but its results deteriorate in terms of sensitivity, especially in Scenario 2. Nonetheless, we note that the `LocFDR` procedure always provides comparable performances to those of the Nollik. The main advantage to the Nollik framework resides in the fully Bayesian approach, which provides straightforward uncertainty quantification and posterior inference. However, the improved performance of our modeling framework comes at a price. As it fully relies on MCMC, Nollik is computationally more expensive than any of the competitors we considered. For example, the average time taken by different Nollik specifications to run ranged between 10 and 30 min to be completed (with the nonparametric specification being the most expensive), while the competitors provided results in a matter of seconds. For a more detailed description of the computational cost of the algorithm, see Section 2.3 of the Supplementary Material.

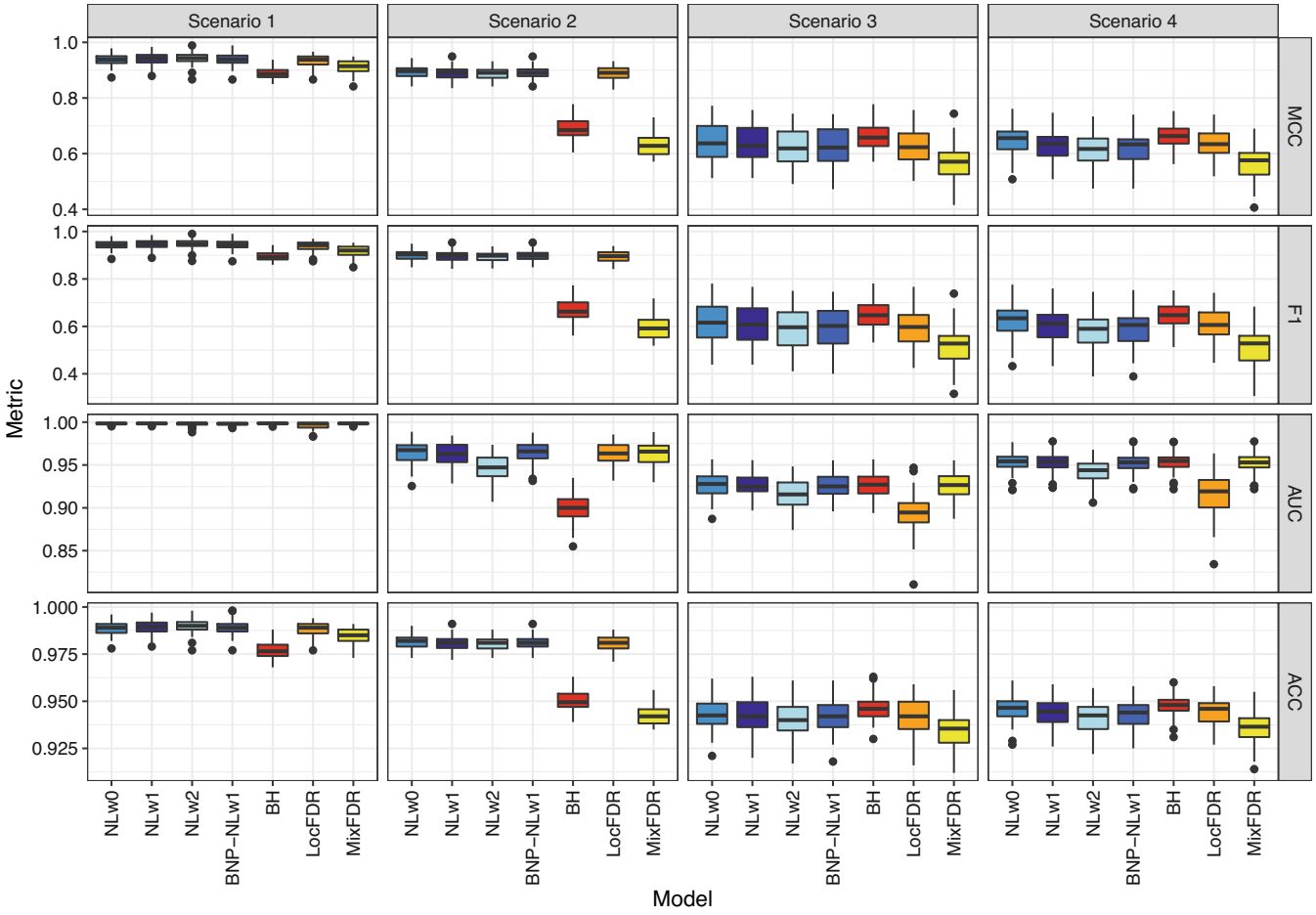

FIGURE 2 Boxplot of the different performance scores (rows) obtained over 50 replications under four scenarios (columns) by the Nollik models (`NLw0-2`) with different weights, the BNP-Nollik model (`BNP-NLw1`), the `BH` procedure, the `LocFDR`, and the `MixFDR`

6 | DIFFERENTIAL GENE EXPRESSION CASE STUDIES

We apply our model to three different gene expression datasets using the weight function w_1 in Equation (15). Results obtained with w_2 are similar; a summary can be found in Section 3.4.2 of the Supplementary Material. We compare the results with Efron's `LocFDR` model and the `BH` procedure. In the first two applications, we run 70 000 MCMC iterations, and, after discarding the first 20 000 as a burn-in period, we thin the remaining chain every 10 iterations. In the third case, we increase the burn-in to 50 000 iterations and thin the chain every 20 steps to reduce autocorrelation. We adopt the hyperparameter configuration detailed in Section 5. On the one hand, we will show how our model can capture the overall data distribution leading to similar results as Efron's `LocFDR`, while also allowing for uncertainty quantification in a coherent, fully Bayesian framework. On the other hand, the mixture formulation allows for more flexible modeling of the irregularities in the empirical null distribution, such as lepto- or platykurtosis (see Figures 10 and 11 in the Supplementary Material). These characteristics are mostly ignored by the `BH` procedure, leading to potential loss of relevant (abundance of irrelevant) genes in the case of leptokurtosis (platykurtosis) of f_0. In Section 3.4.3 of the Supplementary Material, we also report plots that summarize the discoveries obtained by the various methods on the different datasets and their pairwise agreements.

6.1 | HIV microarray data

A benchmark example of gene expression case study is the `HIV` microarray matrix.[1,11] The dataset is publicly available in the R package `locfdr`. The experiment goal is to compare the gene expression values of 7680 microarray genes of 4 HIV

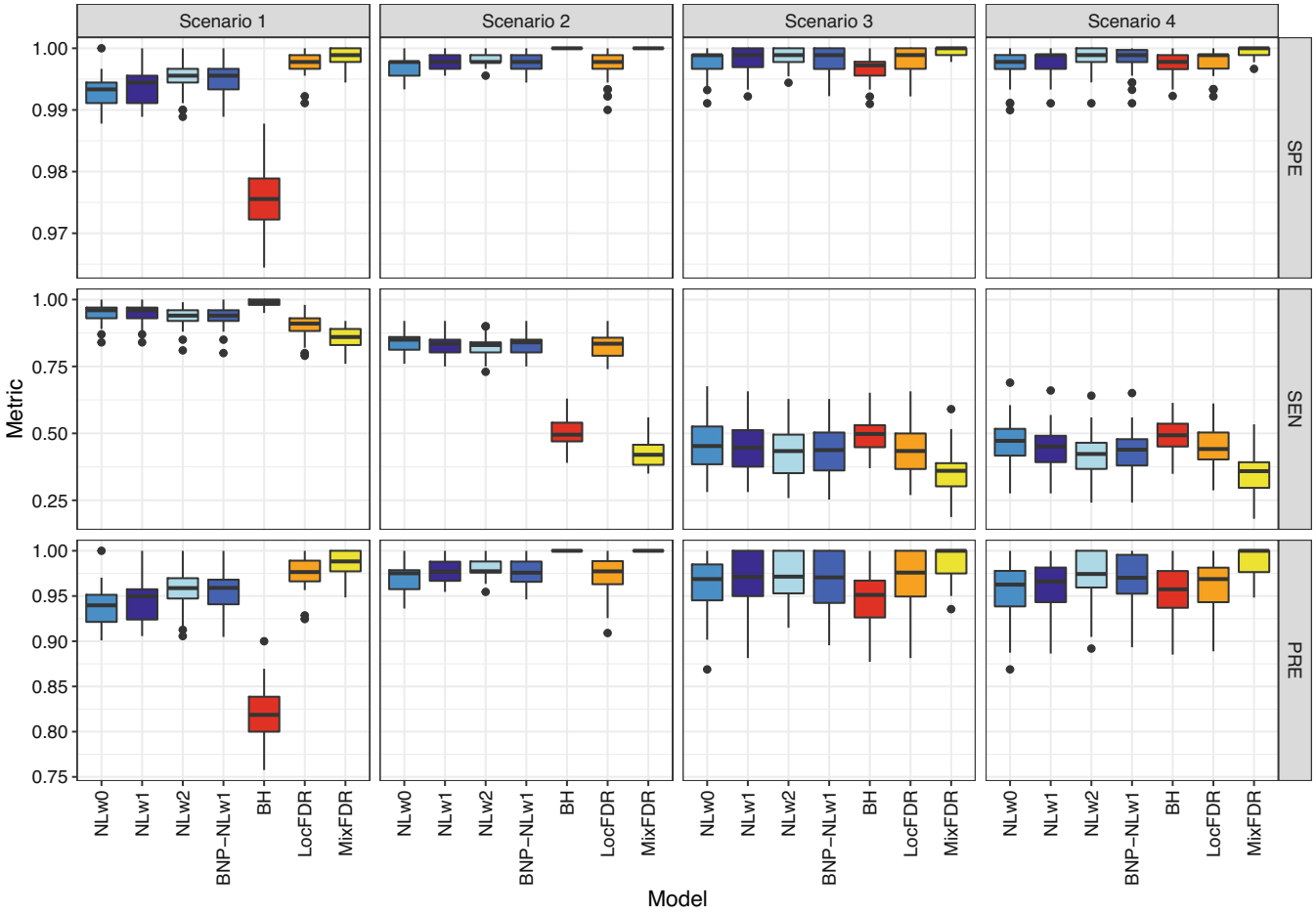

FIGURE 3 Boxplot of the different screening rates (rows) obtained over 50 replications under four scenarios (columns) by the Nollik models (`NLw0-2`) with different weights, the BNP-Nollik model (`BNP-NLw1`), the `BH` procedure, the `LocFDR`, and the `MixFDR`

negative subjects with 4 HIV positive patients. Microarray data are continuous, therefore for each gene we compute the corresponding t-statistics to test the difference of expression among the two groups. We transform the data using the c.d.f. of a Student's t-distribution with 6 degrees of freedom. Efron's `LocFDR` (thresholded at 0.2) flags 160 genes as relevant, the `MixFDR` flags 64 genes, while the `BH` (thresholded at 0.05) only 18. Nollik, in its parametric version, estimates a proportion of relevant hypotheses of $\hat{\rho} = 0.079$ (*s.d.* 0.011), whereas the estimated proportion of the over-expressed genes among the flagged ones is $\hat{\alpha} = 0.121$ (*s.d.* 0.050). The parameter $\hat{\xi}$ of the weight function is estimated as 2.062 (*s.d.* 0.306). The empirical null is characterized by $\hat{\mu}_0 = -0.108$ (*s.d.* 0.012) and $\hat{\sigma}_0^2 = 0.557$ (*s.d.* 0.023), which suggests the potential presence of correlation across tests. We control for a BFDR level of 5%, corresponding to a threshold on $\hat{P}_1(z)$ equal to 0.840. This leads to 143 genes flagged as relevant. From the left panel of Figure 4 we can see how the functions $\hat{P}_1(z)$ for our method and Efron's `LocFDR` are very similar, while the `MixFDR` results in the most conservative method, with the lowest $\hat{P}_1(z)$ on the entire domain.

6.2 | Microbiome abundance table: the Torondel dataset

Many models and software have been developed by bioinformaticians to address the challenges that count data from sequencing studies raise for investigating differential expression (eg, *edgeR* and *baySEQ*[32,33]). Among these models, Love et al[34] have proposed *Deseq2*, a method for differential analysis based on negative binomial regression. To conduct multiple hypothesis testing, *Deseq2* thresholds the `BH` adjusted *P*-values computed from estimated Wald statistics. Here, we apply this method to the `Torondel` dataset,[35] available from the R library `microbiomeSeq`. The abundance table comprises the frequencies of 8883 taxa found in 81 pit latrines: 29 from Tanzania, 52 from Vietnam. Let x_{ij} denote the frequency for

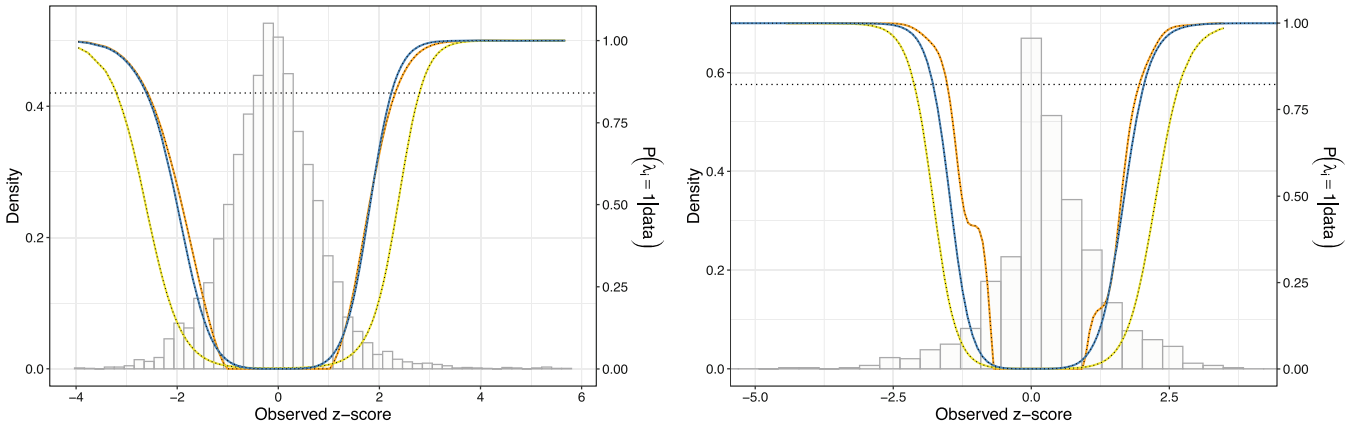

FIGURE 4 `HIV` (left) and `Torondel` (right) datasets. Histograms of the data with function $P_1(z)$ superimposed computed via the `LocFDR` (orange), the `MixFDR` (yellow) and Nollik (blue). The horizontal dotted lines represent the estimated threshold controlling for a BFDR of 5%

taxon i in the pit latrine j. We first filter out all the taxa having variance of the relative counts $r_{ij} = x_{ij}/\sum_{j\geq 1} x_{ij}$ lower than 10^{-7}. The inclusion of this extremely sparse taxa might distort the analysis producing a high number of negligible test statistics which may mislead the estimation of f_0. After these preprocessing steps, we are left with 1204 taxa. The Wald statistics are known to be asymptotically normal in the number of samples. A preliminary data analysis shows that the assumption is reasonable for the rescaled Wald statistics. The `BH` procedures flags only 1 taxon as relevant, the `MixFDR` flags 39 taxa, while the `LocFDR` 107. To better address irregularities in the tails of the data, we employ the BNPNollik. We obtain a proportion of relevant taxa equal to $\hat{\rho} = 0.139$ ($s.d.$ 0.023), while $\hat{\xi} = 2.913$ ($s.d.$ 0.689). Controlling for a BFDR at level 0.05 induces a threshold at 0.822, with 91 taxa marked as relevant, as reported in the right panel of Figure 4, in between the discoveries provided by `MixFDR` and `LocFDR`.

6.3 | Grouped proteomic data: The ubiquitin-protein interactors dataset

In numerous studies, the case group may be composed of J different subsets reflecting specific experimental conditions (eg, stages of a disease, drug dosages, etc.), while the control group remains the same. At one extreme, a separate analysis for each subgroup would result in a potential loss of statistical power. At the other extreme, pooling all the data together is not optimal, since test statistics are not independent across subgroups. In other words, we need to capture commonalities across subgroups induced by the comparisons with a shared control group. To conduct a unified analysis, we extend our approach into a hierarchical model where we jointly estimate different Nollik distributions, one for each experimental condition. In this specification, the dependence across groups induced by the shared control set is captured via a common relevant proportion ρ. Let z_{ij} be the test statistics relative to hypothesis i in the j-th subgroup. Model (2) becomes

$$z_{ij}|\rho, f_{0j}, f_{1j} \sim f_j(z_{ij}) = (1-\rho) f_{0j}(z_{ij}) + \rho f_{1j}(z_{ij}), \tag{17}$$

where f_{0j} and f_{1j} are subgroup-specific null and alternative distribution, respectively. Within each subgroup, the model is (3)-(5), with $\boldsymbol{\theta}_j$ being its specific set of parameters. At the same time, ρ is the same across all the subgroups. The advantages of this model specification are threefold. First, the efficiency and interpretability of the Nollik model are unaltered. Second, the parameters ρ captures the commonality structure, allowing for the borrow of information across subgroups. Third, this model allows the estimation of $P_{1j}(z)$ functions specific for each group, capturing the differences in the various proportions of relevant hypotheses across conditions.

We analyze a mass spectrometry proteomic data for differential protein expression, freely available in the R package `DEP`. The proteins are grouped into three sub-groups, reflecting the different intensities of label-free quantification of the mass spectrometry used to preprocess the data: `Ubi1`, `Ubi4`, and `Ubi6`. See Zhang et al[36] for additional details on the data. We follow the data analysis pipeline indicated by Zhang et al,[37] and obtain 1899 values of proteomic expressions by evaluating the contrasts of the three experimental conditions with the common control group. We evaluate the differential

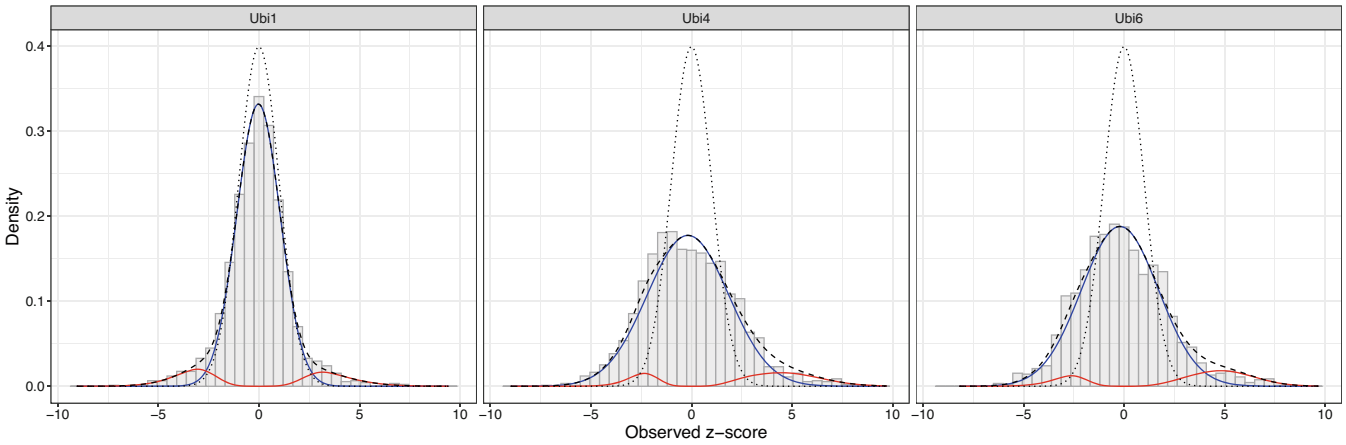

FIGURE 5 `Ubiquitin-protein interactors` dataset. Estimated null (black, dashed), alternative (red), and overall (blue) densities for the of the three subgroups considered (z-scores shown as histograms). The dotted lines denote the $\phi(0, 1)$ densities. Weight function: w_1

TABLE 1 `Ubiquitin-protein interactors`

	Posterior estimates			#Relevant proteins			
	$\hat{\xi}_j$	$\hat{\alpha}_j$	Threshold	Nollik	MixFDR	LocFDR	BH
Ubi1	2.656 (*s.d.* 0.405)	0.447 (*s.d.* 0.249)	0.815	107	61	92	132
Ubi4	2.202 (*s.d.* 0.508)	0.348 (*s.d.* 0.157)	0.903	17	7	13	466
Ubi6	2.366 (*s.d.* 0.539)	0.387 (*s.d.* 0.121)	0.882	31	10	19	457

Note: Estimates and numbers of relevant proteins according to different methodologies stratified by subgroup (`Ubi1`, `Ubi4`, `Ubi6`).

expressions with *Limma*, an Empirical Bayes procedure that produces *moderate t-statistics*, computed as $d/(s + s_0)$, where d is the difference in the sample means, s is the pooled SD and s_0 is a small constant, added to avoid divisions by an extremely small variance estimate.[38]

The estimated overall proportion of relevant tests is $\hat{\rho} = 0.101$ (*s.d.* 0.014). Figure 5 shows the data and the estimated densities stratified by condition. The subgroup-specific models lead to the estimation of different parameters and numbers of relevant proteins, as summarized in Table 1, where we also report the discoveries obtained by the competitors when applied to each subgroup independently. Here, we observe that the platykurtic shape of the histograms `Ubi4` and `Ubi6` lead the `BH` procedure to likely overestimate the number of relevant proteins.

7 | DISCUSSION

In this article, we have proposed a weighted alternative density for multiple hypothesis testing, which leverages on non-local distributions. We have shown how a non-local alternative likelihood can be used as a convenient working density for hypotheses' screening, as it increases the separation from the null distribution. In particular, the trimodal structure of the proposed parametric model, with parameters appropriately tuned for the screening of a large number of hypotheses, allows the segmentation of the z-scores into under-expressed, null, and over-expressed once they are assigned to the normal distributions centered in $\mu_1 < 0$, μ_0, and $\mu_2 > 0$, respectively. In summary, the results we observed in the various simulation studies suggest that Nollik is particularly suited for the following scenarios: in case of an asymmetric behavior of over- and under-expressed test statistics, under potential hyperprior misspecification for the proportion of the relevant statistics, and when the overlap of the two competing distributions hinders the estimation of the empirical null.

An acknowledged limit of our approach stands in its computational cost. Albeit efficient, our implementation fully relies on MCMC, which makes it more costly than every competitor we considered. One solution to this problem would be the development of a variational Bayes approximation[39] of the Nollik posterior, which would scale up its applicability to massive collections of hypotheses.

Methodologically, the simple yet flexible structure of the Nollik models paves the way for relevant extensions. First, the weight functions can be readily generalized to accommodate multivariate data, following Johnson and Rossell.[10] Given a d-dimensional vector $\boldsymbol{z}$, we can define the quantity $Q(\boldsymbol{z}) = \frac{(\boldsymbol{z}-\boldsymbol{z}_0)'\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{z}_0)}{n\xi\sigma^2}$, where Σ is a positive definite matrix and σ^2 and ξ are scalars, and then extend the weight functions to $w_1(\boldsymbol{z};\xi) = 1 - \exp\left[-Q(\boldsymbol{z})^k\right]$ and $w_2(\boldsymbol{z};\xi) = \exp\left[-Q(\boldsymbol{z})^{-k}\right]$. This multivariate likelihood can be useful, for example, in spatial settings, where hypotheses are typically associated with clusters. Second, a covariate-adjusted framework can be naturally addressed without increasing the complexity of the model. Let $\boldsymbol{X}$ be a dataset of P dimensional measurements and denote as $\boldsymbol{X}_i = \left(X_{i1}, \ldots, X_{ip}\right)$ the vector of values specific for individual i. Then we can introduce the dependence via $\boldsymbol{\Lambda}$, specifying $\lambda_i \sim Bern\left(p_i\right)$, $p_i = g(\boldsymbol{X}_i, \boldsymbol{\eta})\ \forall i$. This formulation has two main advantages: (i) the tractability of the MCMC is not altered, being $\boldsymbol{\Lambda}$ separated from the other parameters in the hierarchical structure; (ii) the covariates directly affect parameters driving allocation to latent classes. The function $g(\boldsymbol{X}_i, \boldsymbol{\eta})$ can be assumed as the usual logistic or probit link, for which efficient samplers are readily available.[40,41]

ACKNOWLEDGMENT

In the initial stage of the development of this article, F.D. was supported as a Ph.D. student by University of Milan-Bicocca and Università delle Svizzera italiana. A.M., F.D., and S.P. acknowledge partial support from the SNF grant 200021_200557. Open Access Funding provided by Universita Cattolica del Sacro Cuore within the CRUI-CARE Agreement.

CONFLICT OF INTEREST STATEMENT

There are no financial disclosure to report, and the authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

Software in the form of `R` and `C++` code is available at the Github repository https://github.com/Fradenti/Nollik_2GM. The datasets that support the findings of this study are openly available in the aforementioned `R` packages.

ORCID

Francesco Denti https://orcid.org/0000-0003-2978-4702
Michele Guindani https://orcid.org/0000-0002-6363-9907

REFERENCES

1. van't Wout AB, Lehrman GK, Mikheeva SA, et al. Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4+-T-Cell lines. *J Virol*. 2003;77(2):1392-1402. doi:10.1128/jvi.77.2.1392-1402.2003
2. Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Am Stat Assoc*. 2004;99(465):96-104. doi:10.1198/016214504000000089
3. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*. 2003;19(10):1236-1242. doi:10.1093/bioinformatics/btg148
4. Do KA, Mueller P, Tang F. A nonparametric Bayesian mixture model for gene expression. *J R Stat Soc Ser C*. 2005;54:1-18.
5. Martin R, Tokdar ST. A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics*. 2012;13(3):427-439. doi:10.1093/biostatistics/kxr039
6. Muralidharan O. An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann Appl Stat*. 2010;6(1):422-438. doi:10.1214/09-AOAS276
7. Rao CR. Weighted distributions arising out of methods of ascertainment: what population does a sample represent? In: Atkinson AC, Fienberg SE, eds. *A Celebration of Statistics*. New York, NY: Springer; 1985:543-569. doi:10.1007/978-1-4613-8560-8_24
8. Azzalini A. A class of distributions which includes the normal ones. *Scand J Stat*. 1985;12(2):171-178.
9. O'Hagan A, Leonard T. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*. 1976;63(1):201-203. doi:10.1093/biomet/63.1.201
10. Johnson VE, Rossell D. On the use of non-local prior densities in Bayesian hypothesis tests. *J R Stat Soc Ser B Stat Methodol*. 2010;72(2):143-170. doi:10.1111/j.1467-9868.2009.00730.x
11. Efron B. Size, power and false discovery rates. *Ann Stat*. 2007;35(4):1351-1377. doi:10.1214/009053606000001460
12. Bayarri MJ, Berger J. Robust Bayesian analysis of selection models. *Ann Stat*. 1998;26(2):645-659. doi:10.1214/aos/1028144852
13. Ruggeri F, Sánchez-Sánchez M, Sordo MÁ, Suárez-Llorens A. On a new class of multivariate prior distributions: theory and application in reliability. *Bayesian Anal*. 2021;16(1):31-60. doi:10.1214/19-BA1191
14. Petralia F, Rao V, Dunson DB. Repulsive mixtures. *Adv Neural Inf Process Syst*. 2012;3:1889-1897.
15. Rossell D, Telesca D, Johnson VE. High-dimensional Bayesian classifiers using non-local priors. In: Giudici P, Ingrassia S, Vichi M, eds. *Statistical Models for Data Analysis*. Studies in Classification, Data Analysis, and Knowledge Organization. Heidelberg: Springer; 2013:305-313. doi:10.1007/978-3-319-00032-9-35

16. Rossell D, Telesca D. Nonlocal priors for high-dimensional estimation. *J Am Stat Assoc*. 2017;112(517):254-265. doi:10.1080/01621459.2015.1130634
17. Efron B. Correlation and large-scale simultaneous significance testing. *J Am Stat Assoc*. 2007;102(477):93-103. doi:10.1198/016214506000001211
18. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann Stat*. 1973;1(2):209-230. doi:10.1214/aos/1176342360
19. Sethuraman AJ. A constructive definition of Dirichlet priors. *Stat Sin*. 1994;4:639-650.
20. Sun P, Kim I, Lee KA. Dual-semiparametric regression using weighted Dirichlet process mixture. *Comput Stat Data Anal*. 2018;117:162-181. doi:10.1016/j.csda.2017.08.005
21. Zellner A. On assessing prior distributions and Bayesian regression analysis with g prior distributions. In: Goel PK, Goel PK, Zellner A, eds. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Amsterdam: North-Holland; 1986:389-399.
22. Dunson DB, Pillai N, Park JH. Bayesian density regression. *J R Stat Soc Ser B Stat Methodol*. 2007;69(2):163-183. doi:10.1111/j.1467-9868.2007.00582.x
23. MacEachern SN. Dependent Dirichlet processes. Technical report. Department of Statistics, The Ohio State Univ; 2000.
24. Denti F, Guindani M, Leisen F, Lijoi A, Wadsworth WD, Vannucci M. Two-group Poisson-Dirichlet mixtures for multiple testing. *Biometrics*. 2020;77:622-633. doi:10.1111/biom.13314
25. Roberts GO, Rosenthal JS. Examples of adaptive MCMC. *J Comput Graph Stat*. 2009;18(2):349-367. doi:10.1198/jcgs.2009.06134
26. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. *J Am Stat Assoc*. 2001;96(453):161-173. doi:10.1198/016214501750332758
27. Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. 2004;5(2):155-176. doi:10.1093/biostatistics/5.2.155
28. Rodríguez A, Dunson DB, Gelfand AE. The nested Dirichlet process. *J Am Stat Assoc*. 2008;103(483):1131-1144. doi:10.1198/016214508000000553
29. Efron B. Microarrays, empirical Bayes and the two-groups model. *Stat Sci*. 2008;23(1):45-47. doi:10.1214/08-sts236rej
30. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R News*. 2006;6(1):7-11.
31. Benajmini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol*. 1995;57(1):289-300. doi:10.2307/2346101
32. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform*. 2010;11:422. doi:10.1186/1471-2105-11-422
33. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007;23(21):2881-2887. doi:10.1093/bioinformatics/btm453
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8
35. Torondel B, Ensink JH, Gundogdu O, et al. Assessment of the influence of intrinsic environmental and geographical factors on the bacterial ecology of pit latrines. *Microb Biotechnol*. 2016;9(2):209-223. doi:10.1111/1751-7915.12334
36. Zhang X, Smits AH, Tilburg GB, et al. An interaction landscape of ubiquitin signaling. *Mol Cell*. 2017;65(5):941-955.e8. doi:10.1016/j.molcel.2017.01.004
37. Zhang X, Smits AH, van Tilburg GB, Ovaa H, Huber W, Vermeulen M. Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nat Protoc*. 2018;13:530-550.
38. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
39. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. *J Am Stat Assoc*. 2017;112(518):859-877. doi:10.1080/01621459.2017.1285773
40. Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Pólya-Gamma latent variables. *J Am Stat Assoc*. 2013;108(504):1339-1349. doi:10.1080/01621459.2013.829001
41. Durante D. Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*. 2019;106:765-779. doi:10.1093/biomet/asz034

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Denti F, Peluso S, Guindani M, Mira A. Multiple hypothesis screening using mixtures of non-local distributions with applications to genomic studies. *Statistics in Medicine*. 2023;42(12):1931-1945. doi: 10.1002/sim.9705