WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

High-Dimensional Bayesian Semiparametric Models for Small Samples: A Principled Approach to the Analysis of Cytokine Expression Data

Giovanni Poli[1] | Raffaele Argiento[2,3] | Amedeo Amedei[4] | Francesco C. Stingo[1]

[1]Department of Statistics, Computer Science, Applications "G. Parenti", Università degli Studi di Firenze, Firenze, Italy | [2]Department of Economics, Università degli Studi di Bergamo, Bergamo, Italy | [3]Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milano, Italy | [4]Department of Experimental and Clinical Medicine, Università degli Studi di Firenze, Firenze, Italy

Correspondence: Francesco C. Stingo (francescoclaudio.stingo@unifi.it)

Received: 29 November 2023 | **Revised:** 26 June 2024 | **Accepted:** 4 July 2024

Funding: This study was supported by the Dipartimenti Eccellenti 2023-2027 ministerial funds (Italy), Ministero dell'Università e della Ricerca, and Italian Ministry of University and Research (Tuscany Health Ecosystem (THE) grant, the MUR-PRIN grant 2022 SMNNKY, CUP B53D23009470006, and MUR-PRIN project 2022CLTYP4).

Keywords: Crohn's disease | cytokines | Dirichlet Process | semiparametric Bayesian modeling

ABSTRACT

In laboratory medicine, due to the lack of sample availability and resources, measurements of many quantities of interest are commonly collected over a few samples, making statistical inference particularly challenging. In this context, several hypotheses can be tested, and studies are not often powered accordingly. We present a semiparametric Bayesian approach to effectively test multiple hypotheses applied to an experiment that aims to identify cytokines involved in Crohn's disease (CD) infection that may be ongoing in multiple tissues. We assume that the positive correlation commonly observed between cytokines is caused by latent groups of effects, which in turn result from a common cause. These clusters are effectively modeled through a Dirichlet Process (DP) that is one of the most popular choices as nonparametric prior in Bayesian statistics and has been proven to be a powerful tool for model-based clustering. We use a spike–slab distribution as the base measure of the DP. The nonparametric part has been included in an additive model whose parametric component is a Bayesian hierarchical model. We include simulations that empirically demonstrate the effectiveness of the proposed testing procedure in settings that mimic our application's sample size and data structure. Our CD data analysis shows strong evidence of a cytokine gradient in the external intestinal tissue.

1 | Introduction

Cytokines are small proteins released by different cells, especially immune cells, essential for coordinating immune responses and cell-to-cell communication. The study of cytokines in the etiology of a range of diseases, particularly in the case of inflammatory bowel diseases (IBD), has gained more attention in recent years due to their role in immune response, inflammation, and tissue morphogenesis (see Monastero and Pentyala 2017; Andoh et al. 2008; Guan and Zhan 2017; Friedrich, Pohin, and Powrie 2019). It is common for studies aimed to understand the role of cytokines in intestinal inflammations to measure the level of several cytokines (e.g., Russo et al. 2021, 2022, Niccolai et al. 2021). However, small sample sizes are not rare in this field of research; consequently, standard approaches based on testing of multiple hypotheses are not powerful enough to produce reliable and useful findings. Due to the experimental design, these limits, combined with issues specific to the technology used to measure cytokine levels, such as censorship or high individual variability, hinder standard statistical methods to produce useful results. Multiple testing procedures, standard in this research field, have a high type-I error risk that can hardly be reduced via *p*-value correction methods due to the weak signal-to-noise ratio that makes them low in power.

Here, we propose an approach to test hypotheses in these contexts effectively. We devise a semiparametric Bayesian model that shares information among homogeneous cytokines. In particular, our proposal combines the information from different measurements to reduce inferential errors. Hence, the model yields a reliable indication for future research thanks to a strong regularization based on clustering. The Bayesian approach offers advantages in this context; for instance, posterior estimates are regularized via information sharing across different groups of cytokines. Previous works, for example, Chekouo et al. (2020), highlighted that modeling approaches based on clustering can lead to robust estimation and reliable statistical inference, even for studies with very small sample sizes. Clustering the effects improves the information shared by the model's components, increasing the sample size that (*may*) contributes to the parameters' marginal distributions and reducing the posterior distribution variance, applying an even stronger regularization with respect to hierarchical structures. Extremely small sample sizes require prespecifying the latent group's behaviors, identifying in the a priori model patterns of interest relevant for the researcher (e.g., a positive vs. an adverse reaction to a treatment or a v-shaped trajectory vs. a flat progression). In this work, we define the two patterns of interest a priori, namely, the null versus nonnull effect, and employ a nonparametric model to detect shared patterns across cytokines. The nonparametric effects allow the model to learn the latent structures of interest and its regularization level from the empirical evidence, that is, from the data. In a Bayesian context, this is possible by assuming a Dirichlet Process (DP, Ferguson 1973) as a sampling model for the nonparametric effects. A realization of the DP is an almost surely discrete random probability measure; consequently, the proposed approach allows the model to identify ties between the parameters that quantify the effects of interest and to return latent clusters of homogeneous effects. Clustering regularization suits the analysis of cytokines well. Cytokines can be redundant in their activity, as they can play similar functions; their production can be related, and one cytokine can stimulate its target cells to increase other cytokines' production (Zhang and An 2007). These considerations indicate that similar (or equal) reaction patterns in two different cytokines can realistically increase the statistical credibility of posterior inference and justify using DP as a nonparametric prior on the effects of interest. The latter approach is common in biomedical applications, particularly in genomics, for tasks such as inferring differential gene expression and variable selection (e.g., Do, Müller, and Tang 2005; Guindani et al. 2014; Dahl, Kim, and Vannucci 2009; Barcella et al. 2016). In these applications, the DP is often used jointly with a spike–slab distribution to test sharp null hypotheses (Canale et al. 2017, 2023). We point out that our approach is not only useful for investigating the data's latent clustering structure, but additionally, since it organizes data in homogeneous clusters with effects of similar magnitude and direction, it returns regularized estimates, a relevant feature in settings with a high-dimensional parameter space and low simple size (see, for instance, MacLehose et al. 2007). Note that more flexible variations of DPs models exist in the literature (e.g., Teh et al. 2004), however, in this paper dealing with small sample sizes, we focus on simpler models as learning complex structures (as the one induced by a hierarchical DP prior) may not be feasible. Finally, the nonparametric model is combined with a fully parametric Bayesian model to adapt to the structure of the experiment, maximizing the information shared through a hierarchical structure for the parameters.

The proposed method outperforms competing methods in simulation studies, and our analysis of the Crohn's disease (CD) data identifies a large set of cytokines that show increased activity in both layers of the inflamed human CD mucosa tissues; commonly employed statistical approaches could not identify any deferentially expressed cytokines. Section 2 describes the experiment and data structure. The proposed probabilistic model is presented in Section 3, and the companion Markov chain Monte Carlo (MCMC) algorithm is introduced in Section 4. In Section 5, we conduct a simulation with sample sizes and data structures similar to our data. Our approach is compared with the Wilcoxon rank test, commonly used in practice, and other model-based approaches. In Section 6, we analyze the data, while in Section 7, the analysis results are summarized, and criticisms are discussed.

2 | Motivating CD Study

This work is motivated by a CD study conducted by the Department of Experimental and Clinical Medicine, University of Florence, and the IBD Unit, Careggi University Hospital in Florence. This joint research effort aims to determine the role of cytokines in CD. Crohn's disease is a type of IBD that mainly affects the gastrointestinal tract with extraintestinal manifestations. The exact cause of CD remains unknown. Diet and stress were thought to be among the main drivers of the disease's etiology, but now researchers know that these factors may aggravate but not cause CD (Tomasello et al. 2016; Adolph et al. 2022). According to modern hypotheses, CD may be caused by our immune response triggered by cytokines themselves.

Our experiment explores the relationship between cytokines and CD, focusing on identifying cytokines involved in the infection and which intestinal tissue layer the infection manifests the most. For this purpose, intestinal samples of both healthy and inflamed tissue were obtained during the surgery for each subject, and each of these tissues was divided into layers where the measurements were collected for each cytokine. In detail, 12 CD patients were recruited. All tissue samples were divided into mucosa, submucosa, and serosa, and the profile of a prespecified set of cytokines was analyzed so that for each patient, six sets of cytokines expressions were recorded. For eight CD subjects, 27 cytokines levels were measured, while for the remaining four CD patients, only a subset of six cytokines were measured; this is due to the cost of the kits used to measure cytokine levels. Cytokine concentrations are expressed in ρ g / mL for all subjects. Venous blood samples were collected in vacutainers for serum separation and centrifuged at 2000 rpm for 10 min at room temperature. Serum was immediately collected and stored at −20° C until the analysis, without being thawed and refrozen. We used specifically assembled MixMatch Human kits with a Luminex MAGPIX detection system (Affymetrix, Thermo Fisher, Vienna, Austria) and followed the manufacturer's instructions. The Lower and Upper Limits of Quantification (LLOQ and ULOQ) for the cytokines and chemokines are reported in the Supporting Information Section C.

Standard statistical approaches for analyzing this data set would consist of evaluating $27 \cdot 3$ different hypotheses based on a maximum of 12 paired data points; moreover, the analysis should account for censored observations. Consequently, the standard approach cannot satisfactorily answer the scientific questions of interest due to the limited information for each cytokine. These considerations motivated the development of a new statistical method for the joint analysis of all cytokines.

3 | Bayesian Semiparametric Model

We introduce a Bayesian semiparametric model that can identify cytokines associated with the inflammatory status. The proposed model takes into account the experimental design and effectively borrows strength across cytokines when suggested by the observed data.

Let y_{jlsi} be the natural logarithm of the expression level of cytokine $j = \{1, \dots, J = 27\}$ in tissue type $l = \{1, \dots, L = 3\}$, that is, mucosa, submucosa, or serosa, for subject $i \in \{1, \dots, n_j\}$, where n_j is the number of subjects for which the level of the cytokine j was measured (i.e., $n_j = 8 \ \forall j \in \{1, \dots, 21\}$ and $n_j = 12 \ \forall j \in \{22, \dots, 27\}$). The index s takes the value 1 for healthy tissues and 2 for inflamed tissues. We assume the following model:

$$y_{jlsi} = \mu_{jl} + \delta_i + \theta_{jl} \cdot \mathbb{I}(s = 2) + \varepsilon_{jlsi} \text{ with } \varepsilon_{jlsi} \mid \sigma_j^2 \sim \mathcal{N}\left(0, \sigma_j^2\right) \tag{1}$$

with parameters μ_{jl} representing the expected level for each cytokine in a specific tissue type, δ_i being a subject-specific random intercept, and θ_{jl} being the main focus of our inference, that is, the tissue-specific expected change in expression between healthy and inflamed tissues. The function $\mathbb{I}(s = 2) = 1$ if $s = 2$ and 0 otherwise. Note that the log transformation is commonly used to obtain almost symmetrical distributions since cytokines have a high concentration of low values and a few very high values.

This study aims to identify deferentially expressed cytokines; since all cytokines in all tissues could potentially be involved in the infection, as the cytokines are proxies of the body's response to the infection, an increase in the levels measured in the inflamed tissues is expected. Given the small sample size typical of this type of experiment, we want to share information across cytokines with similar reaction patterns, that is, across cytokines that exhibit differences in the average expression level in each of the three tissue types between inflamed and healthy tissues. Mathematically, we want to identify cytokines that share the same value of $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \theta_{j3})^\top$. This goal is achieved via a Bayesian nonparametric approach. In particular, we assume a DP process as a generative model for the $\boldsymbol{\theta}_j$'s and use the induced clustering to identify groups of cytokines behaving similarly in terms of change in expression between healthy and inflamed tissues. So that the resulting clustering has medical relevance and sensible interpretation.

In the following sections, we first present the prior distributions for the parametric part of the model, that is, parameters $\mu_{jl}, \delta_i, \sigma_j^2$ along with the associated hyperparameters, and then for the nonparametric part, that is, parameters θ_{jl} along with the associated hyperparameters. A directed acyclic graph (DAG) representation of the model structure is available in Section A of the online Supporting Information.

3.1 | Prior Distributions: Parametric Model

The hierarchical structure of the parametric component of the model is specified as follows:

$$\begin{aligned} &\mu_{jl} \mid \xi_j, \tau_j \sim \mathcal{N}(\xi_j, \tau_j^2) \qquad \delta_i \mid \lambda \sim \mathcal{N}(0, \lambda^2) \\ &\sigma_j^2 \mid \sigma_0 \sim \mathcal{IG}\left(\frac{k_\sigma}{2}, \frac{k_\sigma \sigma_0^2}{2}\right) \xi_j \sim \mathcal{N}(m_j, s_j^2) \qquad \tau_j \sim \mathcal{U}(0, t_j^{max}) \\ &\lambda \sim \mathcal{U}(0, l^{max}) \qquad \sigma_0 \sim \mathcal{U}(0, s^{max}). \end{aligned} \tag{2}$$

Cytokines measured in the experiment exhibit observed values of different scale, for example, the sample means in our data range from $3.25 \cdot 10^4 \rho$ g / mL for ICAM-1 to 2.33 ρ g / mL for TNF-α. To model this variability, we assume, for a given j, a hierarchical model for the tissue-specific parameters μ_{jl} where ξ_j is the cytokines overall mean and τ_j^2 the variance for the random effects so that information on the same cytokine can be shared across tissues. On the other hand, subject-specific random effects $\delta_i \sim \mathcal{N}(0, \lambda^2)$ allow the model to capture the correlation across measurements taken from the same subject.

For the standard deviation of the random effects, λ and τ_j, in the absence of strong a priori information, following Gelman (2006), we assume a uniform prior. We complete the parametric part of the model with a hierarchical structure for residual variances, assuming $\sigma_j^2 \mid \sigma_0 \sim \mathcal{IG}\left(\frac{k_\sigma}{2}, \frac{k_\sigma \sigma_0^2}{2}\right)$. Different variance parameters account for heteroskedasticity across cytokines. This hierarchical structure shrinks the variance parameters around a common parameter σ_0^2 that represents the prior expected variance of a new cytokine (e.g., not detected by experimental design). The strength of regularization is controlled by the hyperparameter k_σ, whereas, similarly to the other standard deviation parameters, we assume $\sigma_0 \sim \mathcal{U}(0, s^{max})$.

3.2 | Prior Distributions: Nonparametric Model

One of the primary focuses of this work lies in the nonparametric component, which strives to establish a versatile framework to model the main parameters of interest, namely, $\boldsymbol{\theta}_j$. The proposed approach seeks to diminish posterior variability by organizing these parameters into clusters of homogeneous cytokines effects. The DP prior is arguably the most used Bayesian nonparametric statistical tool. One of the reasons for its popularity is its use in model-based clustering. Indeed, a DP realization is an almost sure discrete random probability measure G. So if $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J$

is a random sample from G, that is, $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_J \mid G \overset{iid}{\sim} G$, then, with positive probability, ties are observed among the $\boldsymbol{\theta}_j$'s. So a clustering can be defined on the cytokines indexes $\{1,\dots,J\}$ by assuming that j and j' belong to the same cluster if $\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j'}$. On the other hand, when investigating sparsity phenomena, Bayesian variable selection is often achieved by assuming a two-component mixture prior for the parameter of interest. Such mixtures are referred to as spike and slab priors. In this paper, we combine the two previously mentioned Bayesian tools to cluster and simultaneously test the relationship between cytokines and inflamed tissues. Indeed, we assume that the generative model for the three-dimensional vector $\boldsymbol{\theta}_j = (\theta_{j1},\theta_{j2},\theta_{j3})^\top$, for $j=1,\dots,J$, is a DP whose base measure is a product of three spike and slab priors featuring an atom at zero.

Our approach builds upon Bayesian testing procedures based on a DP whose base measure is a two-component mixture (Guindani, Müller, and Zhang 2009; Do, Müller, and Tang 2005) and it recalls the approach of Dahl, Kim, and Vannucci (2009). Among other works combining Bayesian nonparametric approaches and variable selection in the biostatistics framework, we refer to Dunson, Herring, and Engel (2008), Yang (2012), and Barcella et al. (2016). A quite interesting methodological study on the consequences of assuming spike and slab prior as centering measures in the Bayesian nonparametric approach is offered by Canale et al. (2017); (2023). The latter works show that this modeling approach increases flexibility and reduces the influence of the prior distributions on the probability of being included in the spike. We argue that it also allows for sharing information among all observed measures (i.e., levels of cytokines) without doing within-model variable selection as the atoms values are cluster-specific. Discrete mass-spikes have many attractive properties, both for inference interpretability and from a theoretical perspective (Barbieri and Berger 2004). However, collecting samples from the posterior distribution may become computationally infeasible as the number of repeated measurements (i.e., analyzed cytokine) increases. A viable alternative, among others, as a base measure in similar contexts could be represented by a continuous spike, with a nonlocal prior (Johnson and Rossell 2010, 2012) as slab distribution.

Summarizing, our prior setting for the $\boldsymbol{\theta}_j = (\theta_{j1},\theta_{j2},\theta_{j3})^\top$ parameters, that is, for the expected change in expression between healthy and inflamed tissues, is assigned as follows:

$$\begin{aligned}
&\boldsymbol{\theta}_j \mid G \overset{iid}{\sim} G, \quad j=1,\dots,J\\
&G \mid \boldsymbol{\pi},\boldsymbol{\omega},\boldsymbol{\eta},M \sim \mathcal{DP}(MG_0) \quad \text{with} \quad G_0 = \bigotimes_{l=1}^{3}\left[\pi_l\,\delta_0(\theta_{jl}) + (1-\pi_l)\,N(\theta_{jl}\mid \eta_l,\omega_l^2)\right]\\
&\pi_l \sim \mathcal{B}eta(a_{\pi_l},b_{\pi_l}) \qquad M \sim \mathcal{G}a(a_M,b_M),
\end{aligned} \tag{3}$$

where $\otimes$ is used to indicate the measures product, $\delta_0(\theta)$ is a Dirac delta function assigning probability one to zero (*spike*), and $N(\theta \mid \eta,\omega^2)$ is the Gaussian density with mean η and variance ω^2 (*slab*).

The base measure G_0 is the product of three spikes and slab distribution modeling the tissue-specific cytokine expression level. The latter (i.e., θ_{jl} $l=1,2,3$), thanks to the spike component, is allowed to be unchanged between inflamed and healthy tissue ($\theta_{jl}=0$). The base measure is fully specified via tree parameters $\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{\pi}$, each of which is a three-dimensional vector. For each tissue $l=1,2,3$, the component π_l of $\boldsymbol{\pi}$, represents the prior probability for the effect θ_{jl} of being 0; the components η_l and ω_l^2 of $\boldsymbol{\omega}$ and $\boldsymbol{\eta}$ are the prior expected value and variance for a nonzero effect, respectively. Note that the slab components of our base measure are independent normals to avoid overparameterization of the models. However, we remark that even if this choice implies conditional independence within clusters, the DPM model can capture the dependence structure of the parameter $(\theta_{j1},\theta_{j2},\theta_{j3})^\top$. In fact, by increasing the number of clusters, our model allows for correlation across the same cytokine effect in various tissue types. To increase the exchange of information between the cytokines expression in the same tissue, an additional level of hierarchy is included in the model assuming each $\pi_l \sim \mathcal{B}eta(a_{\pi_l},b_{\pi_l})$. Each pair $(\eta_l;\omega_l^2)$ is fixed via an empirical Bayes strategy discussed in Section 3.3. In summary, the nonparametric approach induces a clustering among cytokines. Each cluster comprises cytokines with the same expression patterns $\boldsymbol{\theta}_j$ across tissues. It is well-known that the estimated clustering is sensible to the choice of the precision parameter M. To robustify posterior inference, then, we assume $M \sim \mathcal{G}a(a_M,b_M)$.

3.3 | Empirical Bayes

Since our nonparametric model is both latent and instrumental, eliciting honest hyperparameter values for prior distribution parameters ω_l and η_l can be challenging. A common practice consists of setting these parameters to summaries of the data. The prior is then data-dependent, and the approach falls under the umbrella of empirical Bayes methods. Even if formally not fully Bayesian, this empirical Bayesian approach is commonly used in the literature, especially in applied works (Efron 2012; van Houwelingen 2014). The theoretical implications, in terms of frequentist properties, of the Empirical Bayesian approach in the context of DP mixture modeling have recently been investigated by Petrone, Rousseau, and Scricciolo (2014) and Donnet et al. (2018). In this framework, Arbel, Corradin, and Nipoti (2021) have proposed an interesting empirical procedure. We follow this latter work to set the hyperparameter η_l and ω_l, $l=1,2,3$ that characterize the slab component of the DPs base measure in our model. From now on, we introduce the notation $\hat{\eta}_l,\hat{\omega}_l^2$ to emphasize that these parameters were set via empirical Bayes. More in detail, for each $l = 1, 2, 3$, we set

$$\hat{\eta}_l = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_j}\left[\sum_{i=1}^{n_j} y_{jl2i} - y_{jl1i}\right]$$

$$\hat{\omega}_l^2 = \frac{1}{J-1}\left(\sum_{j=1}^{J}\frac{1}{n_j}\left[\sum_{i=1}^{n_j} y_{jl2i} - y_{jl1i}\right] - \hat{\eta}_l\right)^2 \cdot 0.25. \quad (4)$$

From a practical point of view, our Empirical Bayesian approach relies on the idea that every parameter θ_{jl} is interpreted as the expected difference between inflamed and healthy tissue. Then, to assign them a prior, we computed the sample mean and variances of these observed differences. The same weight was used for each cytokine. Moreover, the variance hyperparameters ω_j^2 were penalized by 0.25 (equivalent to half the standard deviation) to induce a larger difference a priori between the spike and the slab distribution.

4 | Posterior Inference

The posterior distribution of the parameters of interest is not available in closed form, and we rely on an MCMC algorithm. To sample the parameters $\boldsymbol{\theta}_j$, we use the infinite mixture representation of the DP process. We have to introduce new notations to provide a brief description of the algorithm. We denote with H the number of observed clusters among the cytokines at any given iteration of the chain, and with $\boldsymbol{\theta}_h^*$ the value of the $\boldsymbol{\theta}_j$ parameter of each cytokine belonging to cluster $h = 1, \ldots, H$. Moreover, the algorithm relies on cluster membership indicator ρ_j, with $\rho_j = h$ implying that cytokine j belongs to cluster h. The base measure G_0 is conjugate to the Gaussian/Normal sampling model defined in Equation (1), so we can exploit algorithm 2 by Neal (2000) to design the MCMC sampling scheme for $\boldsymbol{\theta}_h^*$ and ρ_j parameters. On the other hand, to update M, we use the popular auxiliary sampler by Escobar and West (1995). The full conditionals for random effects variances result in a right-truncated inverse-gamma distribution; similarly, σ_0^2 full conditional results in a right-truncated gamma. The sampling scheme uses the auxiliary variables sampler proposed by Damien and Walker (2001) for both steps. We summarize the steps of our MCMC algorithm hereafter.

- Sequentially updates all the clustering membership variables ρ_j (Neal 2000).
- Update all unique values $\boldsymbol{\theta}_h^*$ sampling from their full conditional (Neal 2000).
- Update M following Escobar and West (1995).
- Update λ^2 and each τ_j^2 following Damien and Walker (2001).
- Update each δ_i and each μ_{jl} sampling from resulting normal distributions.
- Update σ_0^2 again following Damien and Walker (2001).

Note that closed-form expressions are available for all full conditionals, aiding the implementation and convergence of the MCMC sampler; the mathematical steps necessary for the calculation of the full-conditional distributions and the pseudocode are available in Online Supporting Information A.

We aim to identify changes in expected expression of inflamed versus healthy tissues, evaluated as $\mathbb{E}[\mathbb{I}(\theta_{jl} = 0) \mid \boldsymbol{y}] = \Pr(\theta_{jl} = 0 \mid \boldsymbol{y})$ and estimated as the corresponding empirical frequency in the MCMC samples for each pair (j, l). Moreover, our inference will focus on the latent partitions (clusters) of cytokines that will be based on the a posteriori similarity matrix; an entrance of the similarity matrix is the posterior probability that cytokines j and j' are in the same cluster, that is, $\mathbb{E}[\mathbb{I}(\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j'} \mid \boldsymbol{y})] = \Pr(\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j'} \mid \boldsymbol{y})$. These posterior probabilities can be estimated by the proportion of MCMC samples in which the two cytokines are allocated to the same cluster. Following the procedures introduced by Wade and Ghahramani (2018), point estimates for latent partition can be obtained by identifying the partition that minimizes the posterior expectation of a given loss function. Usually, the Binder Loss (BL) or the Variation Information Loss (VI) are adopted. Both loss functions are implemented in the R package mcclust.ext (Wade 2015) that only require an estimate of the similarity matrix as input. Note that the event $\theta_{jl} = \theta_{j'l} = 0$ has positive probability for some l even while the other component of $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{j'}$ are different. However, this event still implies the presence of two clusters. Therefore, the magnitude of not null effects differs between the two clusters. Since the defined approach is designed for small sample sizes, we suggest avoiding using a deterministic criterion for statistical inference, and we argue that inference is better if carried out by inspecting the posterior inclusion probabilities and clustering results jointly. However, objective inferential procedures based on deterministic criteria are necessary in many applications. To this purpose, we suggest the criteria proposed in Newton et al. (2004), which define an intuitive procedure based on posterior probability to obtain a list of significantly differently expressed quantities while bounding the rate of false detections.

5 | Simulation Studies

We designed simulation studies to investigate the proposed approach's finite sample properties and compare them to alternative methods. These studies are based on simulated data that mimic the characteristics of the data described in Section 2.

5.1 | Data Simulation Scenarios

We simulate data for $J = 30$ cytokines and $n_j = 10$ subjects $\forall j = 1, \ldots, J$. Data were generated with the following scheme: we set the expected value of $\mathbb{E}[\mu_{jl}] = \tilde{\xi}_j$ using fixed values (see Online Supporting Information Section B) and sampled μ_{jl} from an $\mathcal{U}(\tilde{\xi}_j - 1, \tilde{\xi}_j + 1)$. Similarly, each δ_i was sampled from an $\mathcal{U}(-0.5, 0.5)$ distribution. The other model's parameters, $\boldsymbol{\theta}_j$, ε_{jlsi}, and σ_j^2 have a scenario-specific setting. We focus on four scenarios (detailed below) and, for each of them, test three settings for the standard deviation; specifically, the parameters σ_j are sampled from a uniform distribution with increasing support that implies scenarios of increasing inferential challenge (low, mid, and high). Given these parameters' values, we generate synthetic data from Equation (1). Note that scenarios differ with respect to the distribution of the residuals. Scenarios are summarized in Table 1. For each of the $4 \cdot 3 = 12$ variance-scenario pairs, we generated 100 data sets.

TABLE 1 | The first four boxes show the effects used to define the four scenarios. The other two boxes concern the distributions used to sample the residual errors and the models compared in these simulation studies.

Scenario I	Scenario II	Scenario III
$\theta_1, \dots, \theta_{10} = (1, 1, 0)^\top$	$\theta_1, \dots, \theta_{20} = \left(u_{j1}, u_{j2}, 0\right)^\top$	$\theta_1, \dots, \theta_{20} = \left(u_{j1}, u_{j2}, 0\right)^\top$
$\theta_{11}, \dots, \theta_{20} = (0.5, 0.5, 0)^\top$	$\theta_{21}, \dots, \theta_{30} = (0, 0, 0)^\top$	$\theta_{21}, \dots, \theta_{30} = (0, 0, 0)^\top$
$\theta_{21}, \dots, \theta_{30} = (0, 0, 0)^\top$	with $u_{jl} \sim \mathcal{U}(0.1,\ 1.0)$	with $u_{jl} \sim \mathcal{U}(0.1,\ 1.0)$
$\varepsilon_{jlsi} \sim \mathcal{N}(0, \sigma_j^2)$	$\varepsilon_{jlsi} \sim \mathcal{N}(0, \sigma_j^2)$	$\varepsilon_{jlsi} \sim \mathcal{SN}(0, \sigma_j^2, 0.99)$

Scenario IV	Standard deviations		Competing methods
$\theta_1, \dots, \theta_{30} = (0, 0, 0)^\top$	Low	$\sigma_j \sim \mathcal{U}(0.25,\ 0.75)$	Semiparametric(Unif.)&(FDR)
$\varepsilon_{jlsi} \sim \mathcal{N}(0, \sigma_j^2)$	Mid	$\sigma_j \sim \mathcal{U}(0.75,\ 1.50)$	Hierarchical(Unif.)&(FDR)
	High	$\sigma_j \sim \mathcal{U}(1.50,\ 3.00)$	Limma
			Wilcoxon rank test
	within the scenarios.		

The four scenarios of interest are summarized as follows:

- **Scenario I** mimics the case where there are true clusters of size 10 for the effects of interest. The first two clusters have a positive effect in two out of three tissues, that is, the third cluster does not react to the infection. Given the presence of groups, Scenario I allows us to control the behavior of latent groups, how the latter affects inference, and evaluate models in terms of cluster performance using the simulation truth. As this is the only scenario where real clusters are present, additional simulations were carried out to gain insights into the sensitivity of the clustering with respect to the prior specifications. These additional simulations are available online as Supporting Information Section B.
- **Scenarios II and III** consider the cases in which a subset of biomarkers react similarly but not identically, that is, there are not true clusters of cytokines, to the infection in two tissues. Here, clustering is useful to reconstruct the distribution of the effects more than organize them in separate groups (see Beraha et al. 2022, for a thorough discussion). In these two scenarios, we set the main effects of interest for the first 20 cytokines to $\theta_j = \left(u_{j1}, u_{j2}, 0\right)^\top$ with $u_{j1}, u_{j2} \sim \mathcal{U}(0.1, 1)$. These scenarios are more likely to mimic the data-generating process of the observed data. Scenarios II and III differ in the distribution of the residuals. In Scenarios I, II, and IV, residuals were generated from $\mathcal{N}(0, \sigma_j^2)$, whereas Scenario III residuals follow a skew-normal distribution (Azzalini 1985). The asymmetry of this distribution is closely related to the parameter $\delta = \alpha / \sqrt{1 + \alpha^2}$. We fixed this parameter equal to 0.99 to obtain a highly positive-skewed distribution (Azzalini 2013), and we reparameterized the distribution to have 0 mean and variance σ_j^2.
- **Scenario IV** represents the null scenario, that is, $\theta_1, \dots, \theta_{30} = (0, 0, 0)^\top$, and is used to control the false positive rate under a no-effects scenario.

5.2 | Competing Methods and Hyperparameter Setting

We compare the performances of our semiparametric model with a few alternative approaches. A first competitor is a commonly used approach based on the Wilcoxon rank test. We performed a test for each cytokine-location pair to compare cytokine levels in healthy and inflamed tissues. We also compare our method to limma (Ritchie et al. 2015), a model-based approach very popular in genomics. Section B of the Online Supporting Information provides a detailed description of the implementation of these methods. Finally, we compare the proposed semiparametric approach with its equivalent fully parametric version, which corresponds to the limit case of $M \to \infty$. This latter corresponds to the parametric model where the $\boldsymbol{\theta}_j$ s are assumed i.i.d. from the base measure of the DP, that is, $\boldsymbol{\theta}_j \mid \boldsymbol{\pi}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\omega}} \overset{iid}{\sim} G_0$; the specification of G_0 is given in Equation (3).

For the proposed Bayesian semiparametric approach and its fully parametric version, we selected values of the hyperparameters that lead to weakly informative prior distributions. In particular, we specify the priors on the ξ_j's as $\xi_j \sim \mathcal{N}(m_j = \tilde{\xi}_j, s_j^2 = 4)$, where $\tilde{\xi}_j$ are the same values used as parameters for the uniform that generated the data. The upper limit of the uniform prior on the standard deviation parameters $\tau_j, \lambda, \sigma_0$ are set to $t_1^{max} = \cdots = t_J^{max} = s_0^{max} = l^{max} = 5$, such that the corresponding variances cannot exceed 25. Finally, we set $M \sim \mathcal{G}a(5, 2)$ as the prior distribution for the precision parameter of the DP. A weekly informative prior for σ_j is achieved by fixing k_σ to 5. We explore two prior strategies for the binary indicator on the inflammation status π_l. The first prior strategy, named Unif., uses a uniform prior distribution, that implies $\mathbb{E}[\theta_{jl} = 0] = 0.5$; the second strategy, named FDR (false discovery rate), uses $\pi_l \sim \mathcal{B}eta(1.8, 0.2)$, that implies $\mathbb{E}[\theta_{jl} = 0] = 0.9$. Compared to the first one, the second is more conservative in terms of detection

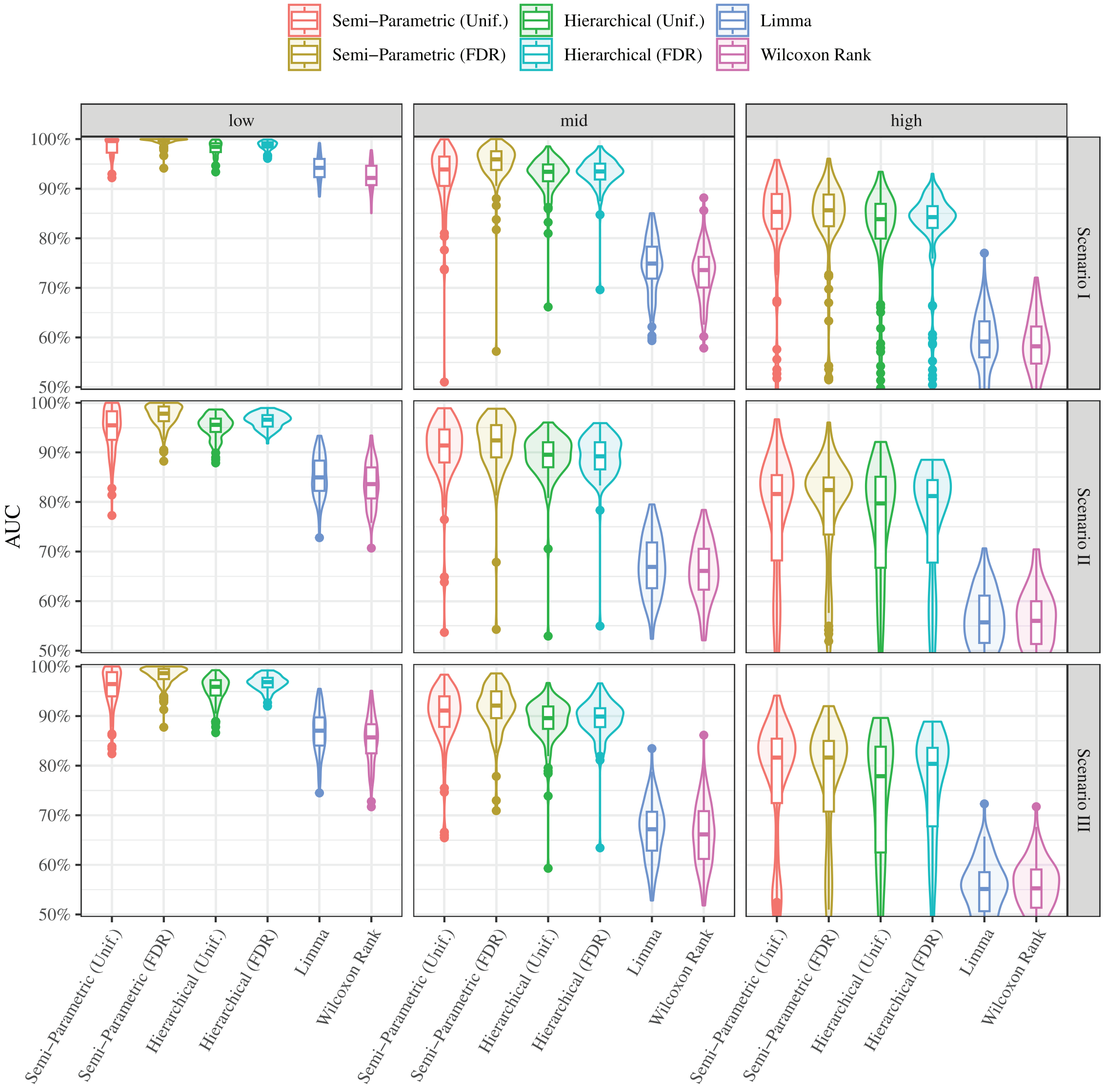

FIGURE 1 | AUC values over 100 replicates for each simulation scenario. The *x*-axis indicates the models and *y*-axis indicates the AUC value. Low, mid, and high labels refer to the support of the standard deviation distribution used to simulate the data.

of differentially expressed cytokines as it reduces a priori the probability of nonzero effects and thus the FDR. The remaining prior parameters are fixed via the Empirical Bayes (EB) procedure described in Section 3.3. For the proposed Bayesian model and its parametric version, we draw 2500 values from the posterior distribution, using a thinning interval of 25 observations after a burn-in period of 100 iterations. The R code used to sample from the posteriors for all models is available on Github.[1]

5.3 | Simulation Results and Discussion

Results of the simulations are summarized in Figures 1 and 2. The main metric on which the models are compared is the AUC calculated using $\Pr(\theta_{jl} = 0 \mid \boldsymbol{y})$ for the Bayesian models and on the *p-values* for the Wilcoxon rank tests test and limma. Given the absence of true positives in Scenario IV, it is impossible to calculate AUC, and the models are compared using the FDR, calculated using several thresholds. To evaluate clustering performances in Scenario I, we obtain point estimates for the data partition, minimizing the Variation of Information criteria and the BL using the approach described in Section 4.

In all scenarios, the Bayesian models perform better than their frequentist counterparts. In particular, semiparametric models always perform slightly better than fully parametric models. For most scenarios, the FDR strategy on the inclusion probabilities has better results than the Unif. strategy. In Scenario III, the results using the two prior strategies are closer. In Scenario IV, regardless of the variances settings, Bayesian models with a conservative prior strategy perform better than the other competing approaches, with the semiparametric approach performing similarly to the parametric one. The evidence favoring the Bayesian approaches is not surprising, given that these models can flexibly capture the dependence structures present in the data and can effectively borrow strength across related cytokines. The

[1] https://github.com/GiovanniPoli/NPB_Cytokines.

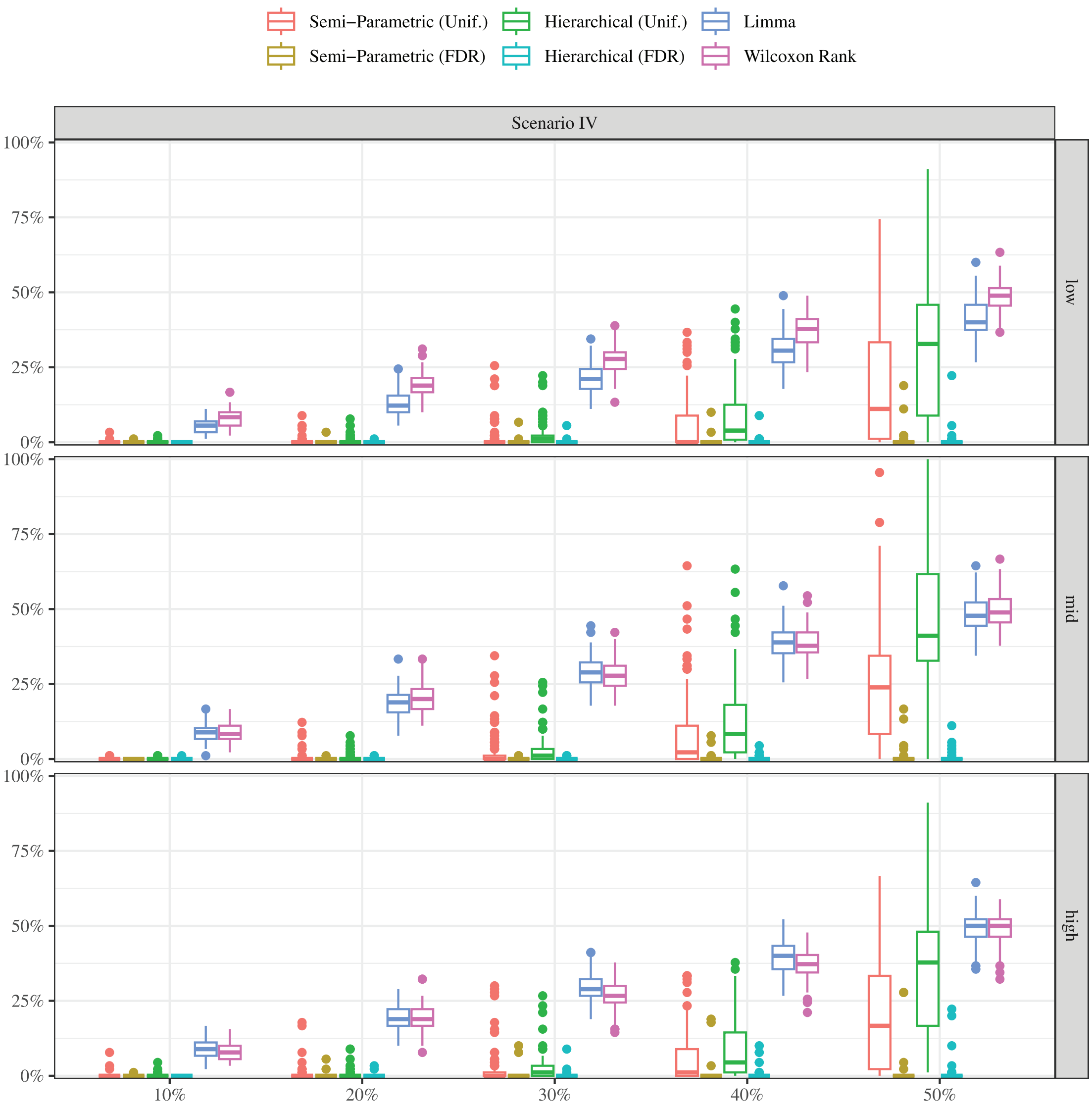

FIGURE 2 | Scenario IV: false-positive rate for the five considered thresholds (10 − 50%). The *x*-axis indicates the considered thresholds, *y*-axis indicates the false-discovery rate (i.e., the number of posterior probabilities above the threshold divided by the total). Low, mid, and high labels refer to the support of the standard deviation distribution used to simulate the data.

main difference between the two Bayesian models lies in the method used to share information and regularize the estimates: in the parametric model, it is driven by the common prior on the parameters $\theta_1, \dots, \theta_J$, whereas in the semiparametric model, information is shared through the cluster structure. Clustering results are summarized in Table 2. The semiparametric model does not always succeed in perfectly reconstructing groups. In all scenarios, for larger values of the error variance, the expected posterior number of clusters (first column) increases; this trend is not necessarily observed if the number of clusters is inferred using point estimation methods such as BL and VI (second and third columns, respectively). This result testifies that perfect reconstruction of the clusters is unnecessary for the semiparametric model to perform well. Furthermore, the difference between the two Bayesian models is more evident in Scenarios II and III than in Scenario I. This may be surprising, considering that the first scenario is the only one with groups of effects. This lack of difference may be due to the larger signal-to-noise ratio of Scenario I with respect to Scenarios II and III. Consequently, the most notable differences are observed in cases with higher error variance. Therefore, the semiparametric model rewards the most when model-based clustering is greatly needed, even when the hierarchical model closely resembles the data-generating mechanism. We conclude that the semiparametric approach is more likely to provide robust inference.

6 | Analysis of the Crohn's Data Set

In this section, we analyze the cytokines data presented in Section 2 using the proposed semiparametric Bayesian model; we use the identical prior setups defined in the simulation studies, with the only exception of the prior on ξ_i, for which the hyperparameters are set based on the upper and lower limits of the cytokines' detection kit used in the experiments; see Online Supporting Information C for more details. As suggested by the

TABLE 2 | Average number of clusters estimated using: (i) the posterior expected number of unique latent parameters (H)—*columns 1–3*; (ii) minimizing the Binder Loss (BL)—*columns 4–6*; or the Variation Information criterion (VI)—*columns 7–9*. Low, mid, and high labels refer to the support of the standard deviation distribution used to simulate the data.

	$\mathbb{E}[\boldsymbol{H} \mid \boldsymbol{y}]$			Clusters via BL			Clusters via VI		
	Low	**Mid**	**High**	**Low**	**Mid**	**High**	**Low**	**Mid**	**High**
				Scenario I					
Semiparametric (Unif.)	3.91	4.04	5.19	4.22	6.99	7.09	2.87	1.64	1.00
Semiparametric (FDR)	3.83	3.97	4.66	4.34	6.97	7.13	2.88	1.75	1.00
				Scenario II					
Semiparametric (Unif.)	4.23	4.65	5.88	5.59	7.39	7.72	2.45	1.17	1.00
Semiparametric (FDR)	4.13	4.61	4.41	5.48	7.70	5.27	2.39	1.18	1.02
				Scenario III					
Semiparametric (Unif.)	4.10	4.69	5.89	5.27	7.54	7.86	2.39	1.21	1.00
Semiparametric (FDR)	4.02	4.67	4.45	5.22	7.82	5.66	2.39	1.24	1.00
				Scenario IV					
Semiparametric (Unif.)	6.57	6.74	6.71	8.00	8.00	8.00	1.00	1.00	1.00
Semiparametric (FDR)	3.20	3.37	3.29	1.20	1.26	1.21	1.00	1.00	1.00

results of the simulation studies, we analyze our data with a semiparametric model using a $Beta(1.8, 0.2)$ as prior distribution for the inclusion probabilities π_l, that is, we use the FDR prior. Four MCMC chains are used to sample from the posterior. Each chain comprises 2500 values sampled after a burn-in period of 500 and a thinning interval of 50. Two different strategies (each twice-repeated) are used to initialize the latent partitions; the first is based on hierarchical clustering, and the second is based on considering each effect as singletons, that is, we start from cytokine-specific effects.

6.1 | Semiparametric Model Inference

Table 3 shows the model inference's main results. Due to high residual variability, the model identifies only one cluster, in which all the cytokines are included with a frequency of around 80% (see top panel of Figure 3). The model captures increasing expression levels of the cytokines in the mucosa and submucosa tissues, with slightly greater confidence in the latter. Parameters related to the serosa tissues, on the contrary, assume a zero value with high probability. In other words, our results remark on an important clinical point, suggesting the inflammation process is essentially localized in the superficial ileal mucosa layers (mucosa and submucosa). This is the key finding of our analysis because, for the first time, we have documented a cytokines gradient in both layers of human CD mucosa tissues.

6.2 | Sensitivity Analysis

We studied the sensitivity of our inference to the specification of the hyperparameters. Specifically, we run the model with different values of the hyperparameters in the prior for M and π_l, of k_σ and with different penalty strategies for the EB variance. In particular, we studied the effect of this hyperparameter choice on

$$\Pr(\theta_{jl} = 0 \mid \boldsymbol{y}),$$

that is, on the posterior probability of the cytokine effect to be null. No significant changes are observed, and all models showed a general behavior similar to the one described in Section 6.1. We mention that some differences in the values of these posterior probabilities are obtained under the parametric hierarchical model, which can be seen as a special case of the nonparametric one when M goes to $+\infty$. However, in this extreme case, the statistical interpretation of the results does not change.

It is well-known that usually, the effect on the posterior estimation of the hyperparameters of the DP parameter M (West 1992; Escobar and West 1995) and the prior on the *inclusion probabilities* π_l (Ishwaran and Rao 2005; Malsiner-Walli and Wagner 2018) can be quite strong. We speculate that this is not the case in our study because the estimated residual variances σ_j^2 are quite high with respect to the estimated effects θ_{jl} (low signal-to-noise ratio). Consequently, the clustering induced by the DP process collapses in a single cluster for any hyperparameter choice. This behavior can be framed into the usual trade-off between density estimation and clustering when setting a DP mixture model. See Beraha et al. (2022), Ghilotti, Beraha, and Guglielmi (2023), and Chandra, Canale, and Dunson (2024) for detailed discussions. To further investigate this behavior, in the next section, we will show how including strong a priori information on the residual variance allows identifying clusters quite interpretable in terms of cytokine effects.

6.3 | Exploratory Analysis on Subgroups

Our analysis of the Crohn's data set has identified a single large group of cytokines. A close look at the similarity matrix

TABLE 3 | Posterior probability for the null effects of the infection (i.e., $\mathbb{E}[\mathbb{I}(\theta_{jl} = 0) \mid \boldsymbol{y}] = \Pr(\theta_{jl} = 0 \mid \boldsymbol{y})$ for each cytokine-tissue pair. Posterior probabilities are estimated by the proportion of MCMC with zero values.

	$\Pr(\theta_{jl} = 0 \mid \boldsymbol{y})$						
	Mucosa	**Submucosa**	**Serosa**		**Mucosa**	**Submucosa**	**Serosa**
GM-CSF	0.0371	0.0187	0.9361	IL-21	0.0468	0.0312	0.9341
sP-Selectin	0.0934	0.1085	0.9251	IL-22	0.0366	0.0179	0.9355
ICAM-1	0.1032	0.0538	0.9297	IL-23	0.0384	0.0175	0.9337
sE-Selectin	0.0069	0.0032	0.9332	IL-27	0.0497	0.0320	0.9337
IFN-α	0.0444	0.0235	0.9342	IL-4	0.0208	0.0065	0.9344
IFN-γ	0.0468	0.0312	0.9317	IL-5	0.0527	0.0249	0.9355
IL-1α	0.0097	0.0046	0.9331	IL-6	0.0207	0.0153	0.9324
IL-1β	0.0094	0.0059	0.9320	IL-8	0.0185	0.0127	0.9317
IL-10	0.0469	0.0238	0.9330	IL-9	0.0409	0.0267	0.9335
IL-12p70	0.0364	0.0151	0.9329	IP-10	0.0286	0.0070	0.9295
IL-13	0.0399	0.0197	0.9347	MCP-1	0.0276	0.0190	0.9308
IL-17A	0.0358	0.0247	0.9336	MIP-1α	0.0206	0.0091	0.9319
IL-18	0.0512	0.0183	0.9342	TNF-α	0.0515	0.0221	0.9315
IL-2	0.0520	0.0304	0.9365				

in Figure 3 (upper panel) suggests that more than one group of smaller effects may be possible. We decided to investigate the nature of this single large group by restricting the residual variance since the regularization encouraged by the model makes it challenging to infer smaller effects. This type of analysis helps identify different patterns that may emerge and then the existence of subgroups. This exploratory analysis is conducted by slightly modifying the semiparametric model by progressively restricting the residual variances; specifically, we replace the hierarchical level for the variances with an informative marginal uniform prior distribution. This approach mimics overconfidence in observed values. In detail, we modify the model such that the $\sigma_j \sim \mathcal{U}(0, s_{max})$, with $s^2_{max} = \{0.5, 0.1\}$, are set to be independent a priori. Additional technical details are presented in Online Supporting Information B.

Results confirm the intuition extrapolated from the analysis of the similarity matrix of the reference model. In Figure 3 (bottom), we note that the single large group identified in the analysis presented in Section 6.1 is split into two subgroups. The first group comprises the majority of cytokines, while the second one is much smaller. Specifically, the second group includes sE-selectin, MIP-1α, MCP-1, IP-10, IL-6, IL-8, IL-1α, and IL-1β that are the cytokines with stronger effects in terms of magnitude and a positive increase also in the serosa layer. The first, larger group behaves similarly to the pattern highlighted by the analysis presented in Section 6.1, with zero coefficients in the serosa layer. sE-selectin is the cytokine of the second group with a pattern closer to the first, larger group. The first group comprises the cytokines that could be the main drivers of the overall inflammatory status of the patients. These results suggest that the second group, comprised of the more proinflammatory cytokines, could drive the inflammation process that extends to the internal ileal layer (serosa). Note, however, that ICAM-1 and sP-selectin are a little further apart from both groups and are the only cytokines that record nonpositive observed sample mean in the mucosa and the ssubmucosa tissues.

7 | Conclusions

In this work, we propose a semiparametric Bayesian approach that improves multiple hypothesis testing by modeling the correlation between cytokines via clustering of the effects of interest. The proposed method leverages two established Bayesian methodologies. On one side, it incorporates a DP prior to induce data clustering. On the other front, it employs a spike and slab prior to assessing the impact of cytokines on inflamed tissues. Beyond the analysis of cytokine data, the proposed approach can be used to test multiple hypotheses in any setting with a positive correlation between repeated measurements. The simulation studies show that for our data structure and sample size, the proposed approach outperforms standard methods used in these contexts, and it is an excellent alternative to other analysis methods.

Our approach offers a framework to get an indication of general patterns of effects by regularizing estimates. While this makes the overall inference reliable in high-dimensional and low-sample-size settings, it may fail to perfectly reconstruct the latent clusters of cytokines when the residual variance is relatively high, as observed in the Crohn's study. We argue that the model applies a strong regularization via larger clusters in contexts where variability does not allow inference on every single effect. In contrast, if specific patterns were evident (minor variance), the clusters would be less populated, and the parameters would be more likely to be interpreted as individual effects. As far as the data analysis is concerned, we found strong evidence in favor

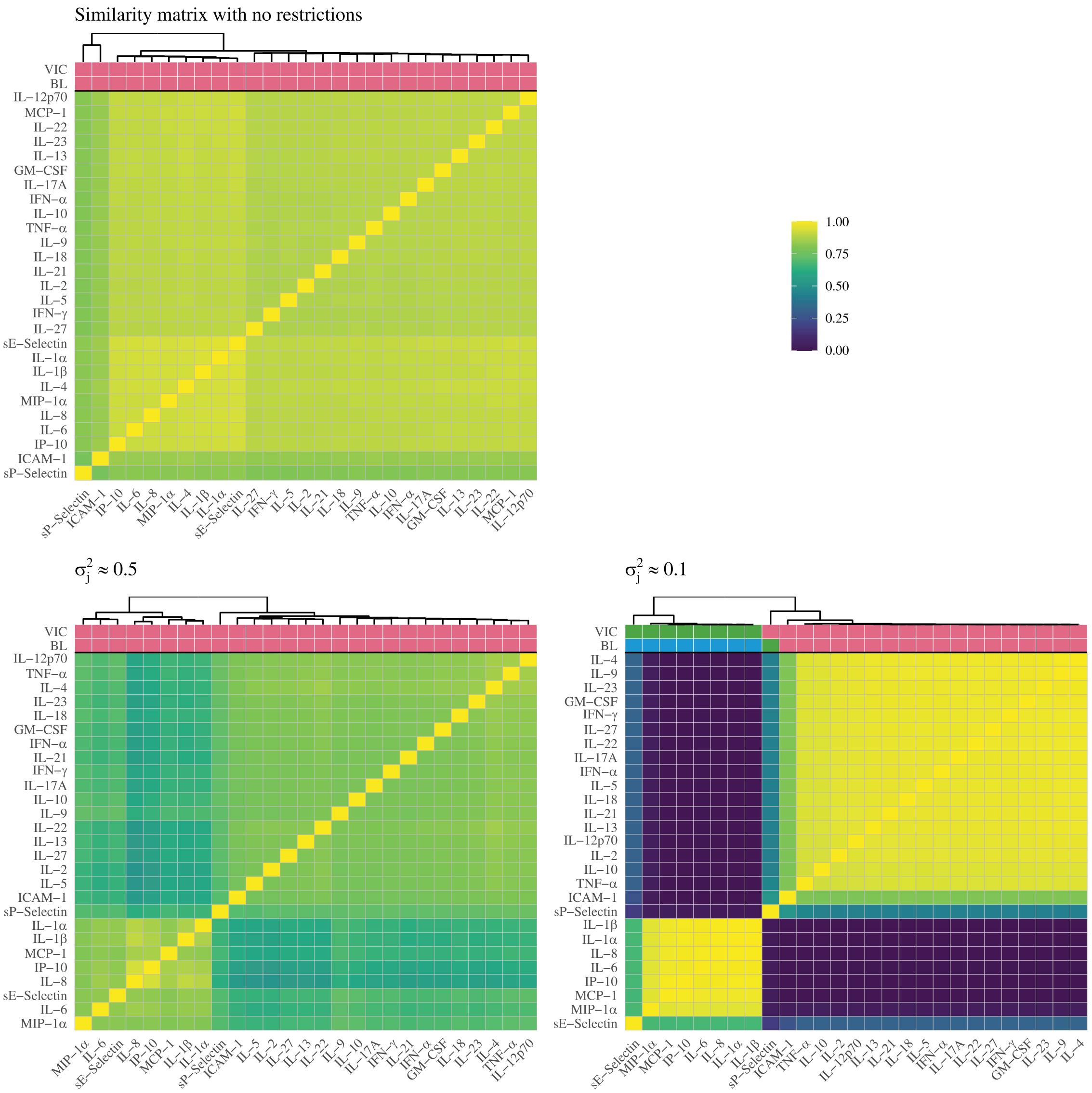

FIGURE 3 | Posterior similarity matrix, represented with heatmaps. Top panel: posterior similarity matrix from the analysis described in Section 6.1. Bottom-left and bottom-right panels: posterior similarity matrix for models with variances restricted using $\sigma_j \sim \mathcal{U}(0, \sqrt{0.5})$ and $\sigma_j \sim \mathcal{U}(0, \sqrt{0.1})$, respectively.

of a general increase in the two outer layers and an indication of a subgroup of cytokines that might represent the core of the immune response to CD that could have a more widespread effect. We also observed that strong a priori information on the residual variances is necessary to obtain meaningful clustering in terms of cytokines effects. This information is needed to contrast the trade-off between density estimation and clustering in Bayesian nonparametric mixture modeling. An alternative solution can be using the repulsive priors mixture model introduced in Beraha et al. (2022) to better distinguish between cytokine effects. Another direction for possible extension is to improve the model regularization over tissues using recent hierarchical extensions of the DPs, such as common atoms models (Denti et al. 2023; Chandra et al. 2023). The approach presented by Chandra et al. (2023) in particular may be pertinent to our case of study because it does not aim to cluster distributions but to cluster similar covariate values across different data sets, sharing information as a consequence. This is part of the current stream of research.

Acknowledgments

The first and last authors were partially supported by the Italian Ministry of University and Research (MUR), Department of Excellence project 2023-2027 ReDS 'Rethinking Data Science' - Department of Statistics, Computer Science, Applications - University of Florence and the MUR-PRIN grant 2022 SMNNKY, CUP B53D23009470006, funded by the European Union Next Generation EU, Mission 4, Component 2. The last author was partially funded by the European Union - NextGenerationEU - National Recovery and Resilience Plan, Mission 4 Component 2 - Investment 1.5 - THE - Tuscany Health Ecosystem - ECS00000017 - CUP B83C22003920001. The second author was partially supported by the project Discrete random structures for Bayesian learning and prediction, funded by the Italian Ministry of University and Research, MUR- PRIN project 2022CLTYP4.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. Data are originated from Russo et al. (2022).

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

References

Adolph, T. E., M. Meyer, J. Schwärzler, L. Mayr, F. Grabherr, and H. Tilg. 2022. "The Metabolic Nature of Inflammatory Bowel Diseases." *Nature Reviews Gastroenterology & Hepatology* 19, no. 12: 753–767.

Andoh, A., Y. Yagi, M. Shioya, A. Nishida, T. Tsujikawa, and Y. Fujiyama. 2008. "Mucosal Cytokine Network in Inflammatory Bowel Disease." *World Journal of Gastroenterology* 14, no. 33: 5154–5161.

Arbel, J., R. Corradin, and B. Nipoti. 2021. "Dirichlet Process Mixtures Under Affine Transformations of the Data." *Computational Statistics* 36, no. 1: 577–601.

Azzalini, A. 1985. "A Class of Distributions Which Includes the Normal Ones." *Scandinavian Journal of Statistics* 12, no. 2: 171–178.

Azzalini, A. 2013. *The Skew-Normal and Related Families*, Vol 3. Cambridge: Cambridge University Press.

Barbieri, M. M., and J. O. Berger. 2004. "Optimal Predictive Model Selection." *The Annals of Statistics* 32, no. 3: 870–897.

Barcella, W., M. D. Iorio, G. Baio, and J. Malone-Lee. 2016. "Variable Selection in Covariate Dependent Random Partition Models: An Application to Urinary Tract Infection." *Statistics in Medicine* 35, no. 8: 1373–1389.

Beraha, M., R. Argiento, J. Møller, and A. Guglielmi. 2022. "MCMC Computations for Bayesian Mixture Models Using Repulsive Point Processes." *Journal of Computational and Graphical Statistics* 31, no. 2: 422–435.

Canale, A., A. Lijoi, B. Nipoti, and I. Prünster. 2023. "Inner Spike and Slab Bayesian Nonparametric Models." *Econometrics and Statistics* 27: 120–135.

Canale, A., A. Lijoi, B. Nipoti, and I. Prünster. 2017. "On the Pitman–Yor Process With Spike and Slab Base Measure." *Biometrika* 104, no. 3: 681–697.

Chandra, N. K., A. Canale, and D. B. Dunson. 2024. "Escaping the Curse of Dimensionality in Bayesian Model-Based Clustering." *Journal of Machine Learning Research* 24, no. 1: 6884–6925.

Chandra, N. K., A. Sarkar, J. F. de Groot, Y. Yuan, and P. Müller. 2023. "Bayesian Nonparametric Common Atoms Regression for Generating Synthetic Controls in Clinical Trials." *Journal of the American Statistical Association* 118, no. 544: 2301–2314.

Chekouo, T., F. C. Stingo, C. A. Class, et al. 2020. "Investigating Protein Patterns in Human Leukemia Cell Line Experiments: A Bayesian Approach for Extremely Small Sample Sizes." *Statistical Methods in Medical Research* 29, no. 4: 1181–1196.

Dahl, D. B., S. Kim, and M. Vannucci. 2009. "Spiked Dirichlet Process Prior for Bayesian Multiple Hypothesis Testing in Random Effects Models." *Bayesian Analysis* 4, no. 4: 707–732.

Damien, P., and S. G. Walker. 2001. "Sampling Truncated Normal, Beta, and Gamma Densities." *Journal of Computational and Graphical Statistics* 10, no. 2: 206–215.

Denti, F., F. Camerlenghi, M. Guindani, and A. Mira. 2023. "A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data." *Journal of the American Statistical Association* 118, no. 541: 405–416.

Do, K.-A., P. Müller, and F. Tang. 2005. "A Bayesian Mixture Model for Differential Gene Expression." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54, no. 3: 627–644.

Donnet, S., V. Rivoirard, J. Rousseau, and C. Scricciolo. 2018. "Posterior Concentration Rates for Empirical Bayes Procedures With Applications to Dirichlet Process Mixtures." *Bernoulli* 24, no. 1: 231–256.

Dunson, D. B., A. H. Herring, and S. M. Engel. 2008. "Bayesian Selection and Clustering of Polymorphisms in Functionally Related Genes." *Journal of the American Statistical Association* 103, no. 482: 534–546.

Efron, B. 2012. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Vol. 1. Cambridge: Cambridge University Press.

Escobar, M. D., and M. West. 1995. "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association* 90, no. 430: 577–588.

Ferguson, T. S. 1973. "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics* 1, no. 2: 209–230.

Friedrich, M., M. Pohin, and F. Powrie. 2019. "Cytokine Networks in the Pathophysiology of Inflammatory Bowel Disease." *Immunity* 50, no. 4: 992–1006.

Gelman, A. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian Analysis* 1, no. 3: 515–534.

Ghilotti, L., M. Beraha, and A. Guglielmi. 2023. "Bayesian Clustering of High-Dimensional Data via Latent Repulsive Mixtures." *arXiv preprint arXiv:2303.02438*.

Guan, Q., and J. Zhan. 2017. "Recent Advances: The Imbalance of Cytokines in the Pathogenesis of Inflammatory Bowel Disease." *Mediators of Inflammation* 2017, no. 6: 1–8.

Guindani, M., P. Müller, and S. Zhang. 2009. "A Bayesian Discovery Procedure." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71, no. 5: 905–925.

Guindani, M., N. Sepúlveda, C. D. Paulino, and P. Müller. 2014. "A Bayesian Semiparametric Approach for the Differential Analysis of Sequence Counts Data." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63, no. 3: 385–404.

Ishwaran, H., and J. S. Rao. 2005. "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies." *Annals of Statistics* 33, no. 2: 730–773.

Johnson, V. E., and D. Rossell. 2010. "On the Use of Non-Local Prior Densities in Bayesian Hypothesis Tests." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72, no. 2: 143–170.

Johnson, V. E., and D. Rossell. 2012. "Bayesian Model Selection in High-Dimensional Settings." *Journal of the American Statistical Association* 107, no. 498: 649–660.

MacLehose, R. F., D. B. Dunson, A. H. Herring, and J. A. Hoppin. 2007. "Bayesian Methods for Highly Correlated Exposure Data." *Epidemiology* 18, no. 2: 199–207.

Malsiner-Walli, G., and H. Wagner. 2018. "Comparing Spike and Slab Priors for Bayesian Variable Selection." *arXiv preprint arXiv:1812.07259*.

Monastero, R. N., and S. Pentyala. 2017. "Cytokines as Biomarkers and Their Respective Clinical Cutoff Levels." *International Journal of Inflammation* 4309485. https://doi.org/10.1155/2017/4309485.

Neal, R. M. 2000. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics* 9, no. 2: 249–265.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist. 2004. "Detecting Differential Gene Expression With a Semiparametric Hierarchical Mixture Method." *Biostatistics* 5, no. 2: 155–176.

Niccolai, E., E. Russo, S. Baldi, et al. 2021. "Significant and Conflicting Correlation of IL-9 With Prevotella and Bacteroides in Human Colorectal Cancer." *Frontiers in Immunology* 11: 573158.

Petrone, S., J. Rousseau, and C. Scricciolo. 2014. "Bayes and Empirical Bayes: Do They Merge?" *Biometrika* 101, no. 2: 285–302.

Ritchie, M. E., B. Phipson, D. Wu, et al. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43, no. 7: e47.

Russo, E., L. Cinci, L. Di Gloria, et al. 2022. "Crohn's Disease Recurrence Updates: First Surgery vs. Surgical Relapse Patients Display Different Profiles of Ileal Microbiota and Systemic Microbial-Associated Inflammatory Factors." *Frontiers in Immunology* 13: 886468.

Russo, E., F. Giudici, F. Ricci, et al. 2021. "Diving Into Inflammation: A Pilot Study Exploring the Dynamics of the Immune–Microbiota Axis in Ileal Tissue Layers of Patients With Crohn's Disease." *Journal of Crohn's and Colitis* 15, no. 9: 1500–1516.

Teh, Y., M. Jordan, M. Beal, and D. Blei. 2004. "Sharing Clusters Among Related Groups: Hierarchical Dirichlet Processes." In *Advances in Neural Information Processing Systems* 17.

Tomasello, G., M. Mazzola, A. Leone, et al. 2016. "Nutrition, Oxidative Stress and Intestinal Dysbiosis: Influence of Diet on gut Microbiota in Inflammatory Bowel Diseases." *Biomedical Papers* 160, no. 4: 461–466.

van Houwelingen, H. C. 2014. "The Role of Empirical Bayes Methodology as a Leading Principle in Modern Medical Statistics." *Biometrical Journal* 56, no. 6: 919–932.

Wade, S. 2015. "Package 'mcclust. ext'." *Journal of Computational and Graphical Statistics* 16: 526–558.

Wade, S., and Z. Ghahramani. 2018. "Bayesian Cluster Analysis: Point Estimation and Credible Balls (With Discussion)." *Bayesian Analysis* 13, no. 2: 559–626.

West, M. 1992. *Hyperparameter Estimation in Dirichlet Process Mixture Models*. Duke University ISDS Discussion Paper# 92-A03.

Yang, M. 2012. "Bayesian Variable Selection for Logistic Mixed Model With Nonparametric Random Effects." *Computational Statistics & Data Analysis* 56, no. 9: 2663–2674.

Zhang, J.-M., and J. An. 2007. "Cytokines, Inflammation and Pain." *International Anesthesiology Clinics* 45, no. 2: 27–37.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.