2019

A Bayesian Nonparametric Multiple Testing Procedure for Comparing Several Treatments Against a Control

Luis Gutiérrez
Pontificia Universidad Catolica de Chile

Andrés Barrientos
Duke University

Jorge González
Pontificia Universidad Catolica de Chile

Daniel Taylor-Rodríguez
Portland State University, dantayrod@pdx.edu

Citation Details

Gutiérrez, Luis; Barrientos, Andrés F.; González, Jorge; Taylor-Rodríguez, Daniel. A Bayesian Nonparametric Multiple Testing Procedure for Comparing Several Treatments Against a Control. Bayesian Anal. 14 (2019), no. 2, 649--675

Bayesian Analysis (2019) **14**, Number 2, pp. 649–675

A Bayesian Nonparametric Multiple Testing Procedure for Comparing Several Treatments Against a Control

Luis Gutiérrez[*], Andrés F. Barrientos[‡],
Jorge González[†], and Daniel Taylor-Rodríguez[§]

Abstract. We propose a Bayesian nonparametric strategy to test for differences between a control group and several treatment regimes. Most of the existing tests for this type of comparison are based on the differences between location parameters. In contrast, our approach identifies differences across the entire distribution, avoids strong modeling assumptions over the distributions for each treatment, and accounts for multiple testing through the prior distribution on the space of hypotheses. The proposal is compared to other commonly used hypothesis testing procedures under simulated scenarios. Two real applications are also analyzed with the proposed methodology.

MSC 2010 subject classifications: Primary 62G10, 62G07; secondary 62G05.

Keywords: Bayes factor, Dependent Dirichlet process, spike and slab priors, shift function.

1 Introduction

We consider the problem of comparing continuous responses for p populations or treatments against those from a control group. This kind of comparisons are in the setting of multiple testing problems, that is, procedures that involve the simultaneous testing of several hypotheses (see e.g., Hochberg and Tamhane, 1987). More formally, let $\mathbf{y} = (\mathbf{y}_c, \mathbf{y}_1, \ldots, \mathbf{y}_p)^t$ be the response variable, where $\mathbf{y}_c = (y_{c,1}, \ldots, y_{c,n_c})^t$ is a random sample of size n_c from the control population, whose distribution G_c is supported on $\mathbb{R}$. Similarly, for $1 \leq k \leq p$, the vector $\mathbf{y}_k = (y_{k,1}, \ldots, y_{k,n_k})^t$ denotes an independent random sample of size n_k from the kth population that is to be compared against the control, and that are distributed according to G_k, also defined on $\mathbb{R}$.

We are interested in finding hypotheses (or equivalently models) that are best supported by the data $\mathbf{y}$ among the set $\mathcal{M} = \{H_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \{0,1\}^p\}$, where $H_{(\gamma_1,\ldots,\gamma_p)}$ represents the model for $\mathbf{y}$ such that

$$\begin{aligned} y_{c,i} &\overset{\text{iid}}{\sim} G_c, i = 1, \ldots, n_c, \\ y_{k,i} &\overset{\text{iid}}{\sim} G_k = \gamma_k G_k^* + (1-\gamma_k) G_c, i = 1, \ldots, n_k \end{aligned} \tag{1}$$

[*]Departamento de Estadística, Pontificia Universidad Católica de Chile, llgutier@mat.uc.cl
[†]Departamento de Estadística, Pontificia Universidad Católica de Chile, jorge.gonzalez@mat.uc.cl
[‡]Duke University, Department of Statistical Science, andres.barrientos@duke.edu
[§]Portland State University, Department of Mathematics & Statistics, dantayrod@pdx.edu

 https://doi.org/10.1214/18-BA1122

$$\text{with } G_c \neq G_k^* \text{ and } k = 1, \ldots, p.$$

For example, with $p = 3$ the hypothesis $H_{(0,0,0)}$ indicates that all treatments are equal to the control, while $H_{(0,1,1)}$ postulates that only the first treatment equals the control. From the Bayesian standpoint, evidence for or against each hypothesis is assessed through their posterior probabilities, and hence all hypotheses considered can be compared against each other, effectively a multiple comparison problem. Specification (1) poses the multiple comparison problem as a model selection one, where the structure and size of the space of hypothesis $\mathcal{M}$ (a.k.a. model space) facilitates a prior specification that accounts for multiple testing. Details on how penalization for model complexity and multiple testing are built into the posterior probabilities is provided in Section 2.

The space of hypotheses $\mathcal{M}$ can be visualized as a partially ordered set through a Hasse diagram (Simovici and Djeraba, 2008). For the case $p = 3$, in Figure 1 each node represents a hypothesis, and the node labels correspond to vectors $\boldsymbol{\gamma} \in \{0, 1\}^p$ defined as above. The edges connecting the nodes indicate nested models represented by the hypotheses. For example, the path in the partially ordered set given by $H_{(0,0,0)} \rightarrow H_{(1,0,0)} \rightarrow H_{(1,0,1)}$, indicates that the model corresponding to the hypothesis where all populations are equal to the control, is nested in the model where only population 1 is different from the control, which is itself nested in the model resulting from the hypothesis where populations 1 and 3 differ from the control. This notion of nesting reflects the way in which model parameters are specified: the parameters of a hypothesis nested within others are also present in the hypotheses that nest it. This nested structure is formalized in Section 3.2. Furthermore, representing the model space as a partially ordered set through this nesting of hypotheses, enables assigning prior probabilities to hypotheses accounting for the model space structure. Again, the Bayesian take on this multiple testing problem consists of identifying, among the 2^p hypotheses in $\mathcal{M}$, those best supported by the data according to their posterior probabilities.

In this paper we have three aims, (I) develop a Bayesian hypothesis testing procedure which yields the posterior probabilities $\pi(H_{\boldsymbol{\gamma}} \mid \mathbf{y})$ for all $H_{\boldsymbol{\gamma}} \in \mathcal{M}$ accounting for multiple testing, (II) relax the strong parametric assumptions in this kind of comparisons and identify differences between groups beyond the location and scale parameters, and (III) if some of the populations are different from the control group, gain understanding about which aspects of their distributions differ.

The comparison of several treatments against a control was first studied by Dunnett (1955) and his proposed method, or extensions of it, are still in use. The method proposed by Dunnett (1955) is based on the construction of confidence statements about the p differences between the mean of each treatment and the mean of the control group. Assuming that the observations are normally distributed, the procedure is based on quantiles of the multivariate t distribution and is capable of testing whether all the differences are simultaneously different from zero with a specified probability. Dunnett and Tamhane (1991) proposed a step-down procedure which provides p-values for the comparisons between the treatments with a control taking into account the multiple comparison nature of the problem. These p-values are also based on the multivariate t distribution. Dunnett and Tamhane (1991) showed that the step-down procedure is

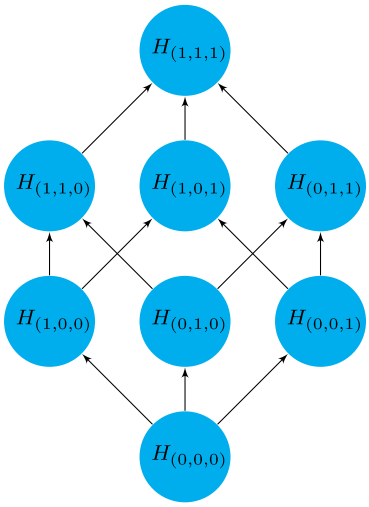


Figure 1: Hasse diagram for the partially ordered set representing the space of hypotheses for three groups against a control.

more powerful than the step-down procedure of Holm (1979) and the single step procedure of Dunnett (1955). The approaches proposed in Dunnett (1955) and Dunnett and Tamhane (1991) assume that observations are normally distributed with unknown treatment means and unknown but common variance. Nonparametric alternatives to the Dunnett test are the Nemenyi-Damico-Wolfe test, which makes one-sided treatment-versus-control multiple comparisons based on joint ranks (Hollander and Wolfe, 1999), and the multiple comparison procedure for unbalanced one-way factorial designs in Gao et al. (2008) based on the unweighted version of the nonparametric relative effects and the associated linear pseudo rank statistics. Both tests are only able to detect possible differences in the locations of the distributions. However, in many applications, the assumptions made by the tests described above are not realistic.

In spite of the lasting popularity of Dunnett's approach, its validity is compromised by its underlying assumptions – normality and equal variances – and its usefulness is limited as comparisons are based only over the location parameters. In fact, distributions may differ in features other than the location, such as tail behavior, symmetry, and number of modes. Although different variances can be accommodated through location-scale families of distributions, the remaining distributional assumptions required (e.g., normality) may still prove to be too stringent to hold in practice. In light of this, and with aims (I) and (II) in mind, we propose a Bayesian nonparametric (BNP) inferential strategy (see, e.g., Ghosh and Ramamoorthi, 2003; Müller and Quintana, 2004; Hjort et al., 2010; Müller and Mitra, 2013), which draws from elements of the Bayesian selection literature (Jeffreys, 1961; Scott and Berger, 2010; George and McCulloch, 1993; Liang et al., 2008; Berger and Pericchi, 1996; Wilson et al., 2010; Womack et al., 2015) to penalize both test multiplicity and the complexity of the underlying models. In addition,

this approach has the advantage of enabling comparisons against the entire distribution of the control group. Aim (III) is addressed by calculating shift functions (Doksum, 1974), which have been extensively used in the literature to test equality of two distributions G_k and G_c (see, e.g., Doksum, 1974; Doksum and Sievers, 1976; Hollander and Korwar, 1980; Wells and Tiwari, 1989; Lu et al., 1994).

BNP approaches for the comparison of two distributions have been proposed in recent years. This kind of comparisons are also known as the two-sample problem. Some contributions are provided in the works of Ma and Wong (2011); Chen and Hanson (2014); Labadi et al. (2014); Holmes et al. (2015); Soriano and Ma (2017). BNP hypothesis testing procedures for two-sample problems in more general spaces such as manifolds are studied in Bhattacharya and Dunson (2012). On the other hand, BNP proposals for multiple testing problems have been developed for specific parameters or applications. For example, Ramanan and Berry (1998) propose a multiple comparison method among the means of p populations based on Dirichlet processes. The work of Scott (2009) describes a framework for multiple hypothesis testing of autoregressive time series, which is used to flag companies whose historical performance is significantly different from that expected due to chance. Kim et al. (2009) use Dirichlet processes with spike and slab priors as a base measure to test jointly for the significance of the random effects in mixed models. Recently, Cipolli et al. (2016) propose a multiple testing procedure for comparing several means based on Polya tree priors. As mentioned before, our approach differs from the existing literature in that it compares multiple populations against a control, identifying differences over the entire distributions. Finally, De Iorio et al. (2004) make a related proposal that uses the Dependent Dirichlet process (MacEachern, 1999, 2000) to define a probability model for random distributions arranged in an analysis of variance (ANOVA)-like array. Although their proposal offers a flexible alternative to model p populations, this model was not formulated as a hypothesis testing procedure.

The remainder of the manuscript is organized as follows: In section 2, the multiple testing problem is formally introduced. In Section 3 we describe the proposed BNP model for hypothesis testing. In Section 4 the proposed model is illustrated using simulated data under different scenarios. This section also provides a Monte Carlo simulation study, where the proposal is compared with current parametric and nonparametric alternatives. Applications using real-life datasets are presented in Section 5. We finalize the paper in Section 6 with conclusions and a discussion.

2 Bayesian testing background

As discussed in the introduction, our goal is to find hypotheses in $\mathcal{M}$ that are best supported by data $\mathbf{y}$. From the Bayesian standpoint this can be done by comparing model posterior probabilities for all models in $\mathcal{M}$. These posterior probabilities are made up of two components: the Bayes Factor (Jeffreys, 1935; Kass and Raftery, 1995) and the prior distribution over $\mathcal{M}$ (Jeffreys, 1961; Scott and Berger, 2006, 2010).

Under hypothesis $H_{\boldsymbol{\gamma}} \in \mathcal{M}$, the likelihood $L(\theta|\mathbf{y}, H_{\boldsymbol{\gamma}})$ connects the data to the parameters θ given hypothesis $H_{\boldsymbol{\gamma}}$. Denote by $\pi(H_{\boldsymbol{\gamma}})$ the prior probability for hypothesis

(model) $H_{\boldsymbol{\gamma}}$, and let $\pi(\theta|H_{\boldsymbol{\gamma}})$ represent the prior density for θ under $H_{\boldsymbol{\gamma}}$. Letting $H_{\mathbf{0}_p}$ represent the hypothesis of all populations being equal to the control, the posterior probability for any $H_{\boldsymbol{\gamma}} \in \mathcal{M}$, can be expressed as:

$$\begin{aligned} \pi(H_{\boldsymbol{\gamma}}|\mathbf{y}) &= \frac{m(\mathbf{y}|H_{\boldsymbol{\gamma}})\pi(H_{\boldsymbol{\gamma}})}{\sum_{H_{\boldsymbol{\gamma}'}\in\mathcal{M}} m(\mathbf{y}|H_{\boldsymbol{\gamma}'})\pi(H_{\boldsymbol{\gamma}'})} \\ &= \frac{B_{\boldsymbol{\gamma},\mathbf{0}_p}(\mathbf{y})\{\pi(H_{\boldsymbol{\gamma}})/\pi(H_{\mathbf{0}_p})\}}{\sum_{H_{\boldsymbol{\gamma}'}\in\mathcal{M}} B_{\boldsymbol{\gamma}',\mathbf{0}_p}(\mathbf{y})\{\pi(H_{\boldsymbol{\gamma}'})/\pi(H_{\mathbf{0}_p})\}}, \end{aligned} \tag{2}$$

where $m(\mathbf{y}|H_{\boldsymbol{\gamma}}) = \int L(\theta|\mathbf{y}, H_{\boldsymbol{\gamma}})\pi(\theta|H_{\gamma})d\theta$ is the marginal likelihood of $\mathbf{y}$ under hypothesis $H_{\boldsymbol{\gamma}}$, obtained by integrating out the model specific parameters from the likelihood. The term, $B_{\boldsymbol{\gamma},\mathbf{0}_p}(\mathbf{y}) = m(\mathbf{y}|H_{\boldsymbol{\gamma}})/m(\mathbf{y}|H_{\mathbf{0}_p})$ is the ratio of marginal likelihoods, also known as the Bayes Factor for hypothesis $H_{\boldsymbol{\gamma}}$ relative to $H_{\mathbf{0}_p}$.

As shown in (2), the model posterior probabilities are specified by the Bayes factor (determined by the priors on the parameter space), and by the priors on the space of hypotheses. The Bayes factor controls for model complexity whereas the prior over the space of hypotheses controls for multiple testing. In the remainder of this section we elaborate on these prior distributions, emphasizing the role that each of them play in the Bayesian testing problem.

2.1 The Bayes factor and the prior over the parameter space

The Bayes factor contains a penalty known as the Ockham's-razor effect (Jeffreys and Berger, 1992). This type of penalization has been commonly associated in the Bayesian variable selection literature as an automatic penalty for model complexity. Although this has not been theoretically established in the nonparametric setting, evidence for this is provided empirically in Basu and Chib (2003). The Bayes factor (and therefore the prior on the parameters), however, does not modulate test multiplicity. This is clear since the Bayes factor between two specific hypotheses (or models) is fixed, regardless of the size of the space of hypotheses (Scott and Berger, 2010).

The strength of the penalty built into the Bayes factor is leveraged by the prior distributions on the model parameters. A considerable body of literature is now available dealing with the definition of these priors for Bayesian testing. Particularly, large efforts have been devoted to the development of "non-informative" priors. Some notable examples are the spike and slab priors (Mitchell and Beauchamp, 1988) and variations of them (see for example George and McCulloch, 1993; Ishwaran and Rao, 2005; Ročková and George, 2014, 2016), g-priors and scale-mixtures of g-priors (Zellner and Siow, 1980; Berger and Pericchi, 1996; Liang et al., 2008; Womack et al., 2014), and non-local priors (Johnson and Rossell, 2010, 2012; Altomare et al., 2013). Given the ease with which spike and slab priors can be adapted, their ability to yield simultaneously selection and estimation, and their time-tested performance, we will adopt a strategy that involves spike and slab components for both location and scale. We provide a detailed development of this prior in Section 3.2.

2.2 Prior distribution over the space of hypotheses

Although the prior probabilities assigned to hypotheses can also penalize model complexity, Jeffreys (1961) recognized that these play an essential role in mediating test multiplicity. The choice of prior distribution on the space of hypotheses requires careful consideration since seemingly innocuous alternatives can have undesirable consequences, and posterior inference is remarkably sensitive to this prior information for small and moderate sample sizes. For example, it has been a common practice to assume equal prior probabilities on all hypotheses; however, this seemingly non-informative alternative favors models of a particular level of complexity, making this choice inadequate.

To define the priors on the space of hypotheses $\mathcal{M}$, each hypothesis is associated to a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \{0,1\}^p$. In Figure 1, these vectors are used to label each node, defining a one-to-one correspondence between the specific configuration of the $\boldsymbol{\gamma}$ vector and a particular hypothesis. As such, the distribution over $\mathcal{M}$ is set in terms of the $\boldsymbol{\gamma}$'s. In comparing a control group against p other populations, $\mathcal{M}$ is populated by the 2^p possible configurations of $\boldsymbol{\gamma}$.

While the importance of these priors was first acknowledged around the 1960's, the literature has only recently gained some attraction. Some examples of priors of this form are the Beta-Binomial(a, b) constructions proposed in (Ley and Steel, 2009; Scott and Berger, 2010; Wilson et al., 2010; Castillo et al., 2015), as well as the construction found in Womack et al. (2015). These priors assume that models of the same size (i.e., hypotheses of the same complexity) obtain the same prior probability. Denoting by $a_{\boldsymbol{\gamma}} = \sum_{k=1}^{p} \gamma_k$ with $a_{\boldsymbol{\gamma}} = l$ (for some $l \in \{0, 1, \ldots, p\}$), this assumption translates into priors for $H_{\boldsymbol{\gamma}}$ of the form

$$\pi(H_{\boldsymbol{\gamma}}) = \pi_p(l) \binom{p}{l}^{-1},$$

where $\pi_p(l) = P(a_{\boldsymbol{\gamma}} = l)$ is the probability for the entire class of hypotheses with exactly l groups differing from the control, so that dividing $\pi_p(l)$ by $\binom{p}{l}$ provides the probability for a single model in the class. As such, priors of this type penalize for the number of hypotheses within the class, as well as for the complexity of models representing the hypotheses in the class.

The prior formulation proposed in Womack et al. (2015) is derived from first principles, has a meaningful and intuitive interpretation, and has been shown to provide suitable penalization for hypothesis complexity as well as test multiplicity. As such, we consider this prior formulation in our model, and defer to Section 3.3 for details on its construction.

3 A BNP model for jointly testing against control

Below we formalize the model specification, providing details on the structure of the likelihood, elaborating on the priors devised for the model parameters, and describing the prior considered for the space of hypotheses. Additionally, for the case when differences have been detected, we briefly discuss the use of shift-functions to identify the zone in the support of the distribution where these differences are found.

3.1 Model definition

Let $y_{k,i}$ be the ith observation in population k, we associate a predictor $x_{k,i}$ indicating the population membership, that is, $x_{k,i} = k$ with $k \in \{c, 1, \ldots, p\}$. We propose the following model,

$$\begin{aligned} y_{k,i} \mid x_{k,i}, \mathcal{P} &\overset{ind}{\sim} \int \phi\left(\cdot|\mu, \sigma^2\right) P_{x_{k,i}}(d\mu, d\sigma^2), \\ \mathcal{P} \mid H_{\boldsymbol{\gamma}} &\sim \pi_{\text{DDP}}(\cdot|H_{\boldsymbol{\gamma}}), \\ H_{\boldsymbol{\gamma}} &\sim \pi_{\mathcal{M}}, \end{aligned} \tag{3}$$

where $\mathcal{P} = \{P_x : x \in \{c, 1, \ldots, p\}\}$, $\phi(\cdot|\mu, \sigma^2)$ is the probability density function of a Gaussian distribution with parameters (μ, σ^2), $\pi_{\text{DDP}}(\cdot \mid H_{\boldsymbol{\gamma}})$ is a prior induced by a Dependent Dirichlet Process DDP (MacEachern, 1999, 2000) under the hypothesis $H_{\boldsymbol{\gamma}}$, and $\pi_{\mathcal{M}}$ is a prior distribution defined on model space $\mathcal{M}$. We now proceed to describe how the priors $\pi_{\text{DDP}}(\cdot|H_{\boldsymbol{\gamma}})$ and $\pi_{\mathcal{M}}$ are defined.

The prior $\pi_{\text{DDP}}(\cdot|H_{\boldsymbol{\gamma}})$ is induced by a process $\mathcal{P}$ whose elements are defined as,

$$P_x(\cdot) = \sum_{j=1}^{\infty} \omega_j \delta_{\left(\mu_j(x), \sigma_j^2(x)\right)}(\cdot), \tag{4}$$

where the weights ω_j are defined with the stick-breaking construction (Sethuraman, 1994), that is, $\omega_j = V_j \prod_{l<j}(1 - V_l)$, with $V_j \mid \kappa \overset{iid}{\sim} \text{Beta}(1, \kappa)$, and δ is the Dirac measure. To provide additional flexibility, we assume that $\kappa \sim \text{Gamma}(a_1, a_2)$. Finally, the atoms are defined as

$$\mu_j(x) = \mu_{c,j} + \eta_{x,j} \quad \text{and} \quad \sigma_j^2(x) = \sigma_{c,j}^2 \tau_{x,j}, \tag{5}$$

where $x \in \{c, 1, \ldots, p\}$. We set $\eta_{c,j} = 0$, and $\tau_{c,j} = 1$, for all j, which is analogous to the standard reference cell constraint. The random sequences $\{\mu_{c,j}\}_{j\geq 1}$ and $\{\sigma_{c,j}^2\}_{j\geq 1}$ are the locations and scales of the control population. Thus, the possible differences detected in the kth population are captured by the changes in locations $\{\eta_{k,j}\}_{j\geq 1}$ and scales $\{\tau_{k,j}\}_{j\geq 1}$ with respect to the control group. When $\eta_{k,j} = 0$ and $\tau_{k,j} = 1$ for a particular k and for every $j \geq 1$, the k-th and control populations are the same.

Remark 1. *Under model (3) we have that* $G_c(\cdot) = \sum_{j=1}^{\infty} \omega_j \Phi(\cdot \mid \mu_{c,j}, \sigma_{c,j}^2)$ *and* $G_k(\cdot) = \sum_{j=1}^{\infty} \omega_j \Phi(\cdot \mid \mu_{c,j} + \eta_{k,j}, \sigma_{c,j}^2 \tau_{k,j})$.

Model (3) belongs to the class of dependent stick-breaking processes, with dependent atoms and common weights. This class enjoys appealing theoretical properties (see for example Barrientos et al., 2012; Pati et al., 2013). The full hierarchical model (3)–(5) bears some similarities to the family of DP mixture models proposed by Müller et al. (2004). Müller et al. (2004) also propose a procedure for a finite number of populations that borrows strength across different but related mixture models. These mixture models relate to each other by their corresponding base measures, which are defined as a mixture of two measures: one idiosyncratic and another common to all populations. Under this

strategy, while the atoms may come from the same distribution these are not exactly the same. This in turns implies that even if two populations have atoms drawn from the same distribution, the mixing distributions for the two populations can differ. Conversely, in our proposal the atoms for two populations can take the same values, hence, the corresponding mixing distributions themselves can be the same. It is the similarity among these mixing distributions what enables us to make comparisons between the control and treatment populations.

3.2 Priors on the atom parameters

In order to test the hypotheses of interest, we require priors for $(\eta_{k,j}, \tau_{k,j})$ that are able to concentrate around 0 for $\eta_{k,j}$ and around 1 for $\tau_{k,j}$ whenever $G_c = G_k$. Here, $k = 1, \ldots, p$ denotes the kth population to be compared with the control group. This type of behavior can be induced using spike and slab priors (George and McCulloch, 1993; Ishwaran and Rao, 2005; Ročková and George, 2016) as the base measure of the Dirichlet process. Thereby, the definition of $\pi_{\text{DDP}}(\cdot|H_{\boldsymbol{\gamma}})$ is completed with the following prior specification,

$$(\mu_{c,j}, \sigma^2_{c,j}) \mid s, b, \epsilon \overset{iid}{\sim} \phi\left(\mu_{c,j} \mid 0, \epsilon s\right) \mathcal{G}\left(1/\sigma^2_{c,j} \mid b/s, b/s\right), \tag{6}$$

$$(\eta_{k,j}, \tau_{k,j}) \mid H_{(\gamma_1,\ldots,\gamma_p)}, s, \epsilon, b \overset{iid}{\sim} \pi_k(\cdot \mid \gamma_k, s, \epsilon, b), \tag{7}$$

where

$$\begin{aligned} \pi_k\left(\cdot \mid \gamma_k, s, \epsilon, b\right) &= \left[\gamma_k \phi\left(\eta_{k,j} \mid 0, \epsilon s\right) + (1-\gamma_k)\phi\left(\eta_{k,j} \mid 0, \epsilon\right)\right] \\ &\quad \times \left[\gamma_k \mathcal{G}\left(1/\tau_{k,j} \mid b/s, b/s\right) + (1-\gamma_k)\mathcal{G}\left(1/\tau_{k,j} \mid b, b\right)\right], \end{aligned}$$

and (s, ϵ, b) are positive hyperparameters with $\mathcal{G}(\cdot|b_1, b_2)$ denoting the density function of the Gamma distribution with mean b_1/b_2 and variance b_1/b_2^2. Notice that ϵ and s control the variance of $\eta_{k,j}$ and $\tau_{k,j}$. As such, ϵ is assumed to be a small near-zero constant and s is fixed in a relative large value. This implies that whenever $\gamma_k = 0$ (i.e., the spike), the $\eta_{k,j}$'s will be close to 0 and the $\tau_{k,j}$'s close to 1, so the atoms for the kth population defined in (5) will concentrate tightly about $\{\mu_{c,j}, \sigma^2_{c,j}\}_{j\geq 1}$. Conversely, if $\gamma_k = 1$, the set of parameters of the kth population are $\{\mu_{c,j}, \eta_{c,j}, \sigma^2_{c,j}, \tau_{k,j}\}_{j\geq 1}$. This spike and slab formulation imposes the nesting structure mentioned in the introduction among hypotheses in $\mathcal{M}$.

In all of our simulations and case studies we fix the hyperparameters used in (6) and (7) at $\epsilon = 0.01$, $s = 1000$ and $b = 100$. For these values, the variance of the slab component ($\gamma_k = 1$) for $\eta_{k,j}$ is 10 and the variance for the spike is 0.01, in both cases the mean of $\eta_{k,j}$ is 0. On the other hand, the spike component for the precision parameter $1/\tau_{k,j}$ has mean 1 and variance 0.01, which implies that the prior induced on the variance parameter $\tau_{k,j}$ is highly concentrated around 1. While, the slab component has mean 1 and variance 10. Thus, in this case $\tau_{k,j}$ can be different from one. These values of the hyperparameters are chosen to calibrate the model for data standardized using the mean and standard deviation for all observations.

3.3 Priors on the space of hypotheses

As mentioned before, we consider the strategy formulated in Womack et al. (2015) to build the priors on the space of hypotheses. These priors control test multiplicity as well as model complexity. Recalling that $a_{\boldsymbol{\gamma}} = \sum_{k=1}^{p} \gamma_k$, this prior formulation assumes that models with equal $a_{\boldsymbol{\gamma}}$ have the same prior probabilities. The prior for the hypothesis that all populations are equal to the control is $\pi(H_{\mathbf{0}_p}) = \rho/(\rho+1)$. The prior for the alternatives hypotheses $H_{\boldsymbol{\gamma}}$ is obtained from the recursion

$$\pi(H_{\boldsymbol{\gamma}}) = \pi_p(l)\binom{p}{l}^{-1} = \left\{\rho \sum_{j=1}^{p-l} \pi_p(l+j)\binom{l+j}{l}\right\}\binom{p}{l}^{-1},$$

where $\boldsymbol{\gamma}$ is a binary vector containing at least one element equal to one, $l = a_{\boldsymbol{\gamma}}$ and $\rho > 0$ is a hyperparameter which fixes the relative odds of belief in a set of local alternatives versus a local null hypothesis.

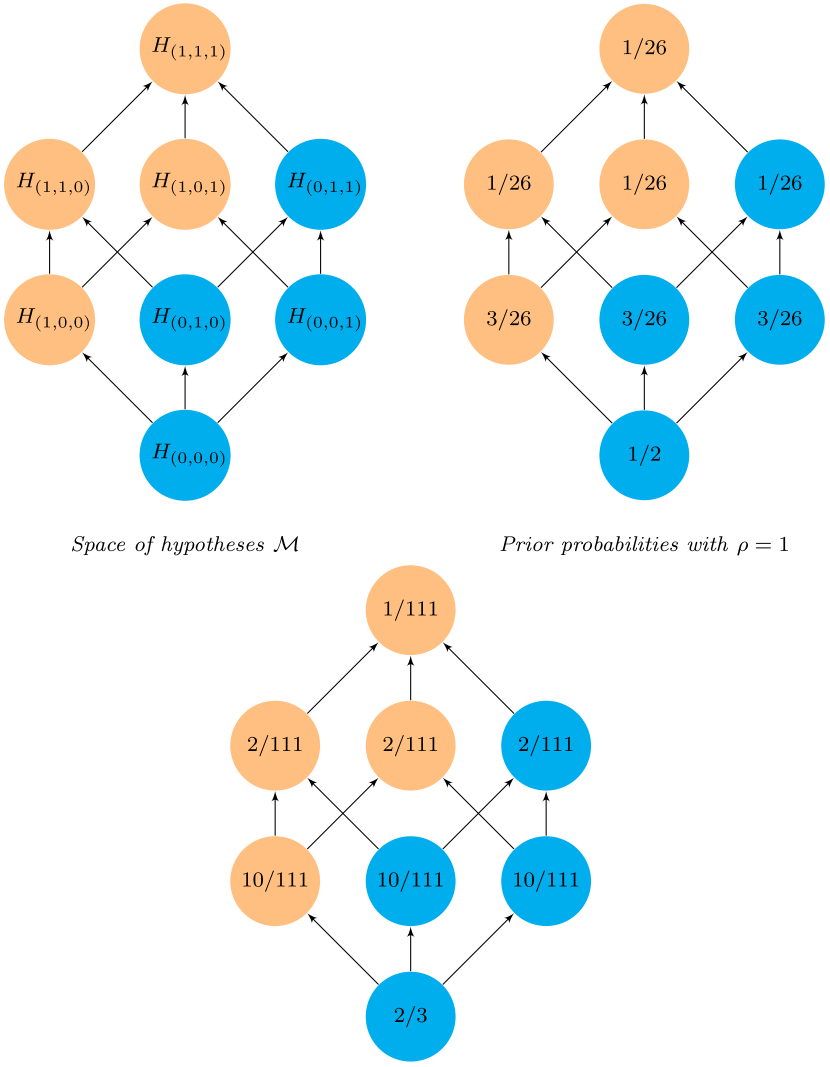
To provide intuition behind this prior construction, consider the example where a control and three additional treatments are to be compared. For instance, focusing on the set of hypotheses $\{H_{(1,0,0)}, H_{(1,1,0)}, H_{(1,0,1)}, H_{(1,1,1)}\}$ (colored orange in Figure 2), note that inside of $\mathcal{M}$, $H_{(1,0,0)}$ is only nested in $H_{(1,1,0)}, H_{(1,0,1)}, H_{(1,1,1)}$. This prior formulation makes the prior probability of $H_{(1,0,0)}$ proportional to the sum of the priors of all hypotheses that nest it, with the proportionality constant given by ρ. For example, for $\rho = 2$, the prior for $H_{(1,0,0)}$ is $10/111 = 2 \times (2/111 + 2/111 + 1/111)$. Turning to hypothesis $H_{(1,1,0)}$, given that it is only nested in $H_{(1,1,1)}$, the prior probability of the former is exactly ρ times the prior probability of the latter.

The same prior structure is observed for any hypothesis in $\mathcal{M}$; it arises from treating each hypothesis as a local null with respect to the set of all hypotheses that nest it (see Womack et al., 2015, for more details on this prior construction). Given that we want to penalize more complex hypothesis but not excessively so, we recommend using $\rho = 1$ as default and make use of this value in the examples and applications of Section 4 and 5.

3.4 Posterior inference and analysis of differences

The posterior inference of model (3) is quite efficient following the slice sampler algorithm described in Kalli et al. (2011) and Walker (2007). This algorithm overcomes the infinite-dimensionality inherent to the Dirichlet process, by considering an augmented model and truncating the number of components in the mixture to a random variable N that takes on finite values drawn as part of the algorithm. Once N is sampled, the posterior inference in each step of the Gibbs algorithm, is reduced to sample the locations, scales, and weights in a finite mixture model. The updating of the locations and scales is facilitated because of the conjugacy in model (3). The weights are simply sampled from a Beta distribution. The posterior probability $\pi(H_{\boldsymbol{\gamma}'} \mid \mathbf{y})$, for $H_{\boldsymbol{\gamma}'} \in \mathcal{M}$ can be approximated with

$$\pi(H_{\boldsymbol{\gamma}'} \mid \mathbf{y}) \approx \frac{1}{B}\sum_{\ell=1}^{B} \mathbf{1}_{\{\boldsymbol{\gamma}^{(\ell)} = \boldsymbol{\gamma}'\}},$$



Space of hypotheses $\mathcal{M}$

Prior probabilities with $\rho = 1$

Prior probabilities with $\rho = 2$

Figure 2: Diagram for the model space for 3 treatments against the control (left), and prior distribution on this space of hypotheses, assuming $\rho = 1$ (center) and $\rho = 2$ (right).

where $\boldsymbol{\gamma}^{(\ell)}$ is the lth posterior sample of $\boldsymbol{\gamma}$, and B denotes the number of posterior samples. Here $\boldsymbol{\gamma}^{(\ell)}$ is updated with a Metropolis-Hastings step. The detail of each step of the Gibbs algorithm is provided in the Supplementary Material (Gutiérrez et al., 2018).

For those groups such that the testing procedure yields evidence for $G_k(\cdot) \neq G_c(\cdot)$ ($k \in \{1, 2, \ldots, p\}$), the *shift function* can be used as a simple alternative to visualize

the aspects in which their distributions differ from the control. Suppose that $G_c(y) \stackrel{d}{=} G_k(y+\Delta_k(y))$, for all y, so that Δ_k can be interpreted as the amount needed to transform the distribution of the control group to the one of the treatments. Doksum (1974) showed that the shift-function, defined as

$$\Delta_k(y) = G_k^{-1}(G_c(y)) - y \tag{8}$$

is the unique function such that the equality in distribution holds.

Thus, $\Delta_k(\cdot)$ characterizes how two independent distributions differ. In fact, if $\Delta_k(y) = 0$, for all y, then the distributions are identical. If $\Delta_k(y) \neq 0$, for some y, the distributions are not equal and we can inspect the set $\{y : \Delta_k(y) \neq 0\}$ to identify where they differ. Equivalently, if treatments are being compared to the control, the set $\{y : \Delta_k(y) \neq 0\}$ gives information on what aspects of the distribution are being influenced by the treatment.

The computation of the shift function is straightforward in our algorithm because for each step of the Gibbs algorithm, we have posterior random realizations of $G_c^{(\ell)}$ and $G_k^{(\ell)}$, $\ell = 1, \ldots, B$. These distributions can be computed using the expressions shown in Remark 1 together with the Gibbs algorithm in the Supplementary Material. Let $G_k^{-1}(u) = \inf\{x : G_k(x) \geq u\}$ be the left inverse of G_k, a random realization of the shift function can be calculated as

$$\Delta_k(y)^{(\ell)} = \begin{cases} G_k^{-1(\ell)}(G_c^{(\ell)}(y)) - y & \text{if } \gamma_k^{(\ell)} = 1, \\ 0, \quad \forall y & \text{if } \gamma_k^{(\ell)} = 0. \end{cases} \tag{9}$$

The realizations in (9) can be used to compute the sample posterior mean $\bar{\Delta}(y)$ as a point estimator of the shift function. Also, a 95% credible set $(\Delta_*(y), \Delta^*(y))$ can be estimated using the 2.5% and 97.5% percentiles of the random realizations of $\Delta_k(y)$. The values of y for which $0 \notin (\Delta_*(y), \Delta^*(y))$ can be used to determine the set $\{y : \Delta_k(y) \neq 0\}$.

4 Illustrations with simulated data and a Monte Carlo study

In this section, we first illustrate the performance of the proposed BNP method using synthetic data. Then, we investigate the performance of the proposed BNP testing procedure using a Monte Carlo simulation study with scenarios that focus on different features that may be of interest in this type of problems. The posterior inference obtained from our method is compared to those from other popular hypothesis testing alternatives.

4.1 Examples with synthetic data

Three examples are provided, each differing in terms of which populations differ from the control, as well as the distributions considered to generate the data (i.e., distributions

that differ in location, scale, asymmetry and the number of modes in the populations). In particular, we consider a skew normal distribution for example 1 with density given by

$$f(x) = \frac{2}{\tau}\phi\left(\frac{x-\mu}{\tau}\right)\Phi\left(\alpha\left(\frac{x-\mu}{\tau}\right)\right),$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ correspond to the density and the cumulative distribution of a standard normal distribution. The parameters μ, τ and α correspond to the location, scale and shape of the distribution, respectively. We denote the skew normal distribution as $\text{SN}(\mu, \tau, \alpha)$. Notice that a skew normal distribution with $\alpha = 0$ is equivalent to a normal distribution. In each example, we consider $p = 3$ populations and a control group with sample sizes $n_c = n_k = 90, \quad k \in \{1, 2, 3\}$. A summary of the settings used for each example is given in Table 1.

Example	True hypothesis	Population			
		$\mathbf{y}_c$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
1	$H_{(0,0,1)}$	$\text{SN}(0,1,0)$	$\text{SN}(0,1,0)$	$\text{SN}(0,1,0)$	$\text{SN}(0,1,1)$
2	$H_{(1,1,0)}$	$\text{N}(0,1)$	$\text{N}(0,2.25)$	$\text{N}(0,0.25)$	$\text{N}(0,1)$
3	$H_{(1,1,1)}$	$\text{N}(0,0.49)$	$\text{N}(0.6,1)$	$0.5\text{N}(-1.2,0.25)+0.5\text{N}(1.2,0.25)$	$0.3\text{N}(-1.2,0.25)+0.7\text{N}(1.2,0.25)$

Table 1: True populations for examples 1 to 3. $\text{N}(\mu, \sigma)$ denotes normal distribution. $\text{SN}(\mu, \tau, \alpha)$ denotes the skew normal distribution with location μ, scale τ and shape α.

The proposed model in (3) was fitted to each of the three data sets generated. The values of the hyperparameters for the total mass parameter κ were fixed at $a_1 = a_2 = 1$, which is a relatively standard choice. We experimented with other choices for a_1 and a_2, for example, $a_1 = 5$, $a_2 = 1$, and in general the posterior quantities were robust to that choices. The posterior value of κ was updated as in Escobar and West (1995), see details in Supplementary Material. A graphical display of the posterior inference showing true vs estimated densities and shift functions for each example, is provided in Figures 3 and 4. The Figures also display the true shift function and its estimation given by the posterior mean and the corresponding credible sets. We also added a red dashed line, which represents $\Delta_k(\cdot) = 0$ as a reference for interpretation purposes. In particular, we display the posterior inference where there are differences between the control and some population. As it can be seen from these figures, the proposed model provides good estimations of the densities and the shift functions. In most cases, estimates follow closely both the true densities and true shift functions and the true model was completely covered by 95% point-wise credible interval. Regarding the posited hypothesis, for each example our testing procedure assigned posterior probability of 1 to the true hypothesis. Thus, in conclusion, the method was able to accurately detect the true hypothesis, estimating correctly both the densities and shift functions.

4.2 Monte Carlo simulation study

The 9 scenarios to be investigated are determined by the levels of two factors: the extent to which the tested populations differ (difference levels which are given by the parameters μ_l, σ_l^2 and θ_l, with $l = 1, 2, 3$, see the details in Table 2), and the sample size ($n = 50, 150, 300$). In all the scenarios, the control population was assumed to be

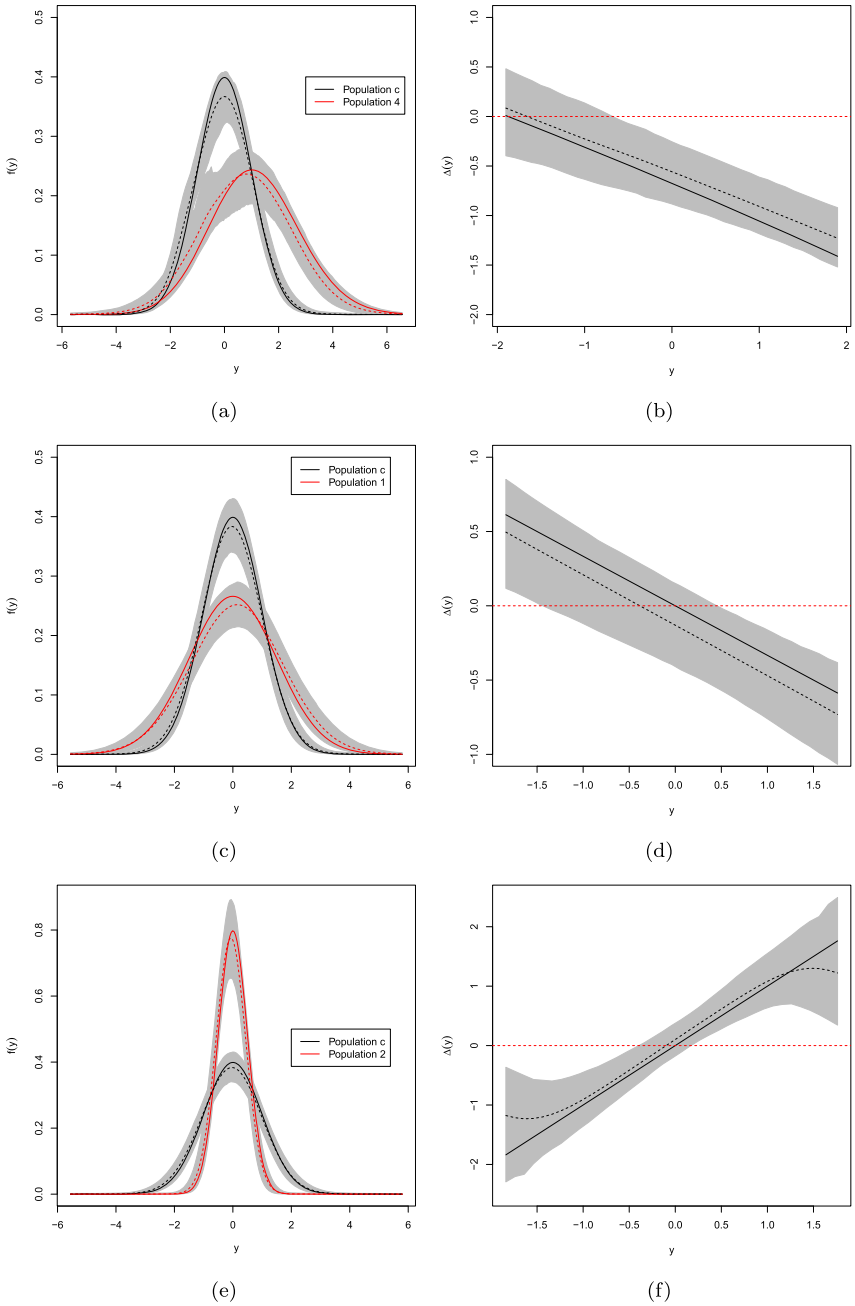


Figure 3: True density (solid lines), posterior mean (dashed lines) and 95% point-wise credible intervals (left panel). True shift function, posterior mean (dashed line) and 95% credible intervals (right panel). Panels (a) and (b) depict Example 1. Panels (c) to (f) correspond to Example 2.

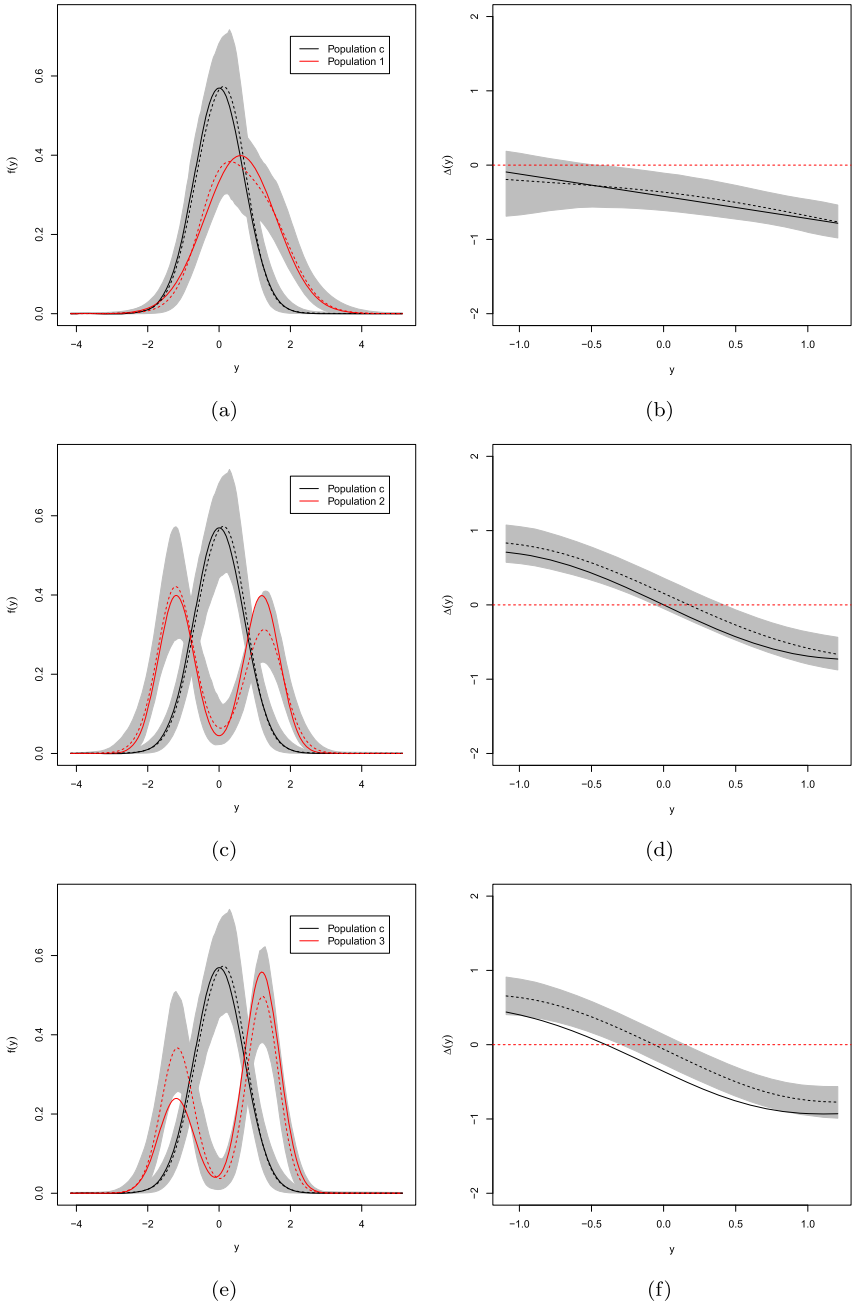


Figure 4: True densities (solid lines), posterior mean (dashed lines) and 95% point-wise credible intervals (left panel). True shift function (solid lines), posterior mean (dashed line) and 95% credible intervals (right panel) right panel. All panels correspond to Example 3.

Difference level	Population	Parameter	True hypothesis							
			$H_{(0,0,0)}$	$H_{(1,0,0)}$	$H_{(0,1,0)}$	$H_{(0,0,1)}$	$H_{(1,1,0)}$	$H_{(1,0,1)}$	$H_{(0,1,1)}$	$H_{(1,1,1)}$
$l=1$	$\mathbf{y}_c \overset{\text{iid}}{\sim} \text{N}(0,1)$	-	-	-	-	-	-	-	-	-
	$\mathbf{y}_1 \overset{\text{iid}}{\sim} \text{N}(\mu_l,1)$	μ_l	0	0.625	0	0	0.625	0.625	0	0.625
	$\mathbf{y}_2 \overset{\text{iid}}{\sim} \text{N}(0,\sigma_l^2)$	σ_l	1	1	1.498	1	1.498	1	1.498	1.498
	$\mathbf{y}_3 \overset{\text{iid}}{\sim} 0.5\text{N}(-\theta_l,1)+0.5\text{N}(\theta_l,1)$	θ_l	0	0	0	1.162	0	1.162	1.162	1.162
$l=2$	$\mathbf{y}_c \overset{\text{iid}}{\sim} \text{N}(0,1)$	-	-	-	-	-	-	-	-	-
	$\mathbf{y}_1 \overset{\text{iid}}{\sim} \text{N}(\mu_l,1)$	μ_l	0	0.819	0	0	0.819	0.819	0	0.819
	$\mathbf{y}_2 \overset{\text{iid}}{\sim} \text{N}(0,\sigma_l^2)$	σ_l	1	1	1.806	1	1.806	1	1.806	1.806
	$\mathbf{y}_3 \overset{\text{iid}}{\sim} 0.5\text{N}(-\theta_l,1)+0.5\text{N}(\theta_l,1)$	θ_l	0	0	0	1.359	0	1.359	1.359	1.359
$l=3$	$\mathbf{y}_c \overset{\text{iid}}{\sim} \text{N}(0,1)$	-	-	-	-	-	-	-	-	-
	$\mathbf{y}_1 \overset{\text{iid}}{\sim} \text{N}(\mu_l,1)$	μ_l	0	1.051	0	0	1.051	1.051	0	1.051
	$\mathbf{y}_2 \overset{\text{iid}}{\sim} \text{N}(0,\sigma_l^2)$	σ_l	1	1	1.994	1	1.994	1	1.994	1.994
	$\mathbf{y}_3 \overset{\text{iid}}{\sim} 0.5\text{N}(-\theta_l,1)+0.5\text{N}(\theta_l,1)$	θ_l	0	0	0	1.652	0	1.652	1.652	1.652

Table 2: Values of the parameters μ_l, σ_l and θ_l, $l \in \{1,2,3\}$.

a standard normal distribution. Population 1 was generated by a normal distribution with variance equal to one and location denoted by μ_l. In the case of population 2, we considered a normal distribution with mean zero and variance denoted by σ_l^2. Finally, population 3 was generated by a 50–50 mixture of two normal distributions. Each component of the mixture has variance equal to one, with locations $-\theta_l$ and θ_l, respectively. Specifications of the populations and the corresponding values for the parameters are shown in Table 2.

The results of the Monte Carlo simulation study are summarized in Figure 4. Each of the 9 scenarios is represented by an image plot. The vertical axis in these plots represent the true models, while the horizontal axis represents the estimated models. Each of the 64 cells shows the average over the 100 Monte Carlo replications for the posterior probability $\pi(H_{\boldsymbol{\gamma}} \mid \mathbf{y})$. A probability value equal to 0 is represented by black, while a value 1 is represented by white in the grayscale. The correctly identified hypotheses are represented in the main diagonal. Figure 4 shows that as the value of l increases (i.e., the difference between the treatments and the control is more evident), the average posterior probability $\pi(H_{\boldsymbol{\gamma}} \mid \mathbf{y})$ approaches 1 in the true model. The same behavior is observed as the sample size increases. We compared the performance of our model against some classical testing procedures. To this end, we select the *maximum a posteriori hypothesis*, given by

$$\hat{H}_{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \quad \pi(\boldsymbol{\gamma}' = \boldsymbol{\gamma} \mid \mathbf{y}).$$

For the classical tests, we consider a significance level of 0.05. In particular, we compared our proposal with the multiple hypothesis testing procedures of Dunnett (Dunnett and Tamhane, 1991), Nemenyi-Damico-Wolfe (Hollander and Wolfe, 1999) and Gao (Gao et al., 2008). We also used some two-sample testing procedures (Welch's t-test, Levene, Wilcoxon and Kolmogorov-Smirnov), and adjust them for multiple comparisons with Bonferroni corrections.

Table 3 provides a summary of the performance of each test. We report the number of times that the true hypotheses, as described in Table 2, were detected for each test. The multiple testing procedures of Dunnett, Nemenyi-Damico-Wolfe and Gao were able to detect differences only in the locations of the distributions, that is, they detect reasonably well the hypothesis H_{100}. The two-sample tests provide the expected results; that is: the t-test and the Wilcoxon test are able to detect differences in locations. Their performances are better than our proposal in detecting this aspect of the distribution, especially in the scenarios with small sample size. The Levene test is able to detect scale differences (H_{010}). This test also detects the hypothesis H_{001}, which could be thought as a scale difference, while actually the difference is due to the mixture specification in Population 3. The performance of Levene test for the scenarios with small sample sizes was better than our proposal. The Kolmogorov-Smirnov test was able to detect differences across the entire distribution. However, the performance of our proposal was better or as good as Kolmogorov-Smirnov in all the scenarios. Image plots showing the number of times that the eight considered test selected each model are provided in Figures S.1 to S.8 in the Supplementary Material.

	$l=1$, $n=50$								$l=2$, $n=50$								$l=3$, $n=50$							
True model	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
H_{000}	**97**	89	92	92	85	84	81	85	**98**	89	92	92	85	84	81	85	**97**	89	92	92	85	84	81	85
H_{100}	32	77	75	73	**82**	2	81	74	66	95	**96**	95	92	2	90	88	95	98	**99**	97	94	2	93	92
H_{010}	43	2	1	1	5	**68**	1	8	82	5	1	1	5	**87**	2	18	**97**	5	1	1	5	92	2	32
H_{001}	46	4	3	1	4	**81**	5	26	81	4	3	2	4	**91**	5	42	**97**	4	5	2	4	93	7	80
H_{110}	8	3	2	3	4	3	6	**12**	**47**	4	3	4	4	4	8	30	**88**	4	4	5	5	4	9	41
H_{101}	8	4	3	3	4	5	6	**16**	**39**	6	5	5	8	5	9	36	**91**	8	6	7	9	5	10	79
H_{011}	7	0	0	0	0	**63**	0	4	48	0	0	0	0	**91**	0	16	83	0	0	0	0	**95**	0	32
H_{111}	**4**	0	0	2	0	0	0	3	**18**	0	0	1	0	1	0	13	**65**	0	0	1	0	1	0	27
	$l=1$, $n=150$								$l=2$, $n=150$								$l=3$, $n=150$							
H_{000}	**98**	94	94	94	86	86	85	83	**100**	94	94	94	86	86	85	83	**98**	94	94	94	86	86	85	83
H_{100}	61	96	**97**	95	91	3	92	94	**97**	96	97	95	91	3	92	94	**100**	96	97	95	91	3	92	94
H_{010}	68	4	1	1	5	**90**	4	38	**100**	5	1	1	5	91	4	71	**100**	5	1	1	5	91	5	88
H_{001}	82	3	1	1	6	**84**	7	74	**99**	4	2	1	7	84	6	97	**100**	5	4	1	5	84	4	99
H_{110}	**44**	4	4	4	4	3	4	34	**98**	4	5	5	4	3	4	85	**100**	4	5	5	4	3	5	95
H_{101}	51	2	2	1	4	6	4	**82**	**96**	3	2	1	3	6	5	94	**100**	2	4	4	1	6	6	96
H_{011}	35	1	1	1	1	**94**	2	45	**94**	1	1	1	1	94	2	83	**100**	1	1	2	1	94	2	97
H_{111}	23	0	0	0	0	5	0	**37**	**89**	0	0	0	0	5	0	88	**99**	0	0	2	0	5	0	96
	$l=1$, $n=300$								$l=2$, $n=300$								$l=3$, $n=300$							
H_{000}	**99**	93	96	96	86	83	86	90	**100**	93	96	96	86	83	86	90	**100**	93	96	96	86	83	86	90
H_{100}	85	**99**	99	98	94	1	94	93	**100**	99	99	98	94	1	94	93	**100**	99	99	98	94	1	94	93
H_{010}	**94**	4	3	2	6	94	7	84	**100**	5	3	3	6	94	7	95	**100**	5	4	3	5	94	7	95
H_{001}	**97**	4	2	1	6	86	9	88	**100**	5	4	1	7	86	10	89	**100**	5	4	2	4	86	8	89
H_{110}	**88**	3	2	2	3	8	4	84	**100**	4	4	3	4	8	6	95	**100**	5	7	6	4	8	6	95
H_{101}	92	3	4	4	3	1	5	**94**	**100**	3	9	8	3	1	3	94	**100**	2	19	18	3	1	3	94
H_{011}	84	0	0	0	1	**98**	1	82	**100**	0	0	0	1	98	1	97	**100**	0	0	0	0	98	1	97
H_{111}	79	0	0	1	0	5	0	**95**	**100**	0	0	3	0	5	0	100	**100**	0	1	7	0	5	0	100

Table 3: Number of times that the true model was detected in the 100 replications. [1]: BNP, [2]: Dunnet, [3]: Nemenyi, [4]: Gao, [5]: t-test, [6]: Levene, [7]: Wilcoxon, [8]: Kolmogorov-Smirnov. Values in boldface indicate the test with the best performance for each hypothesis.

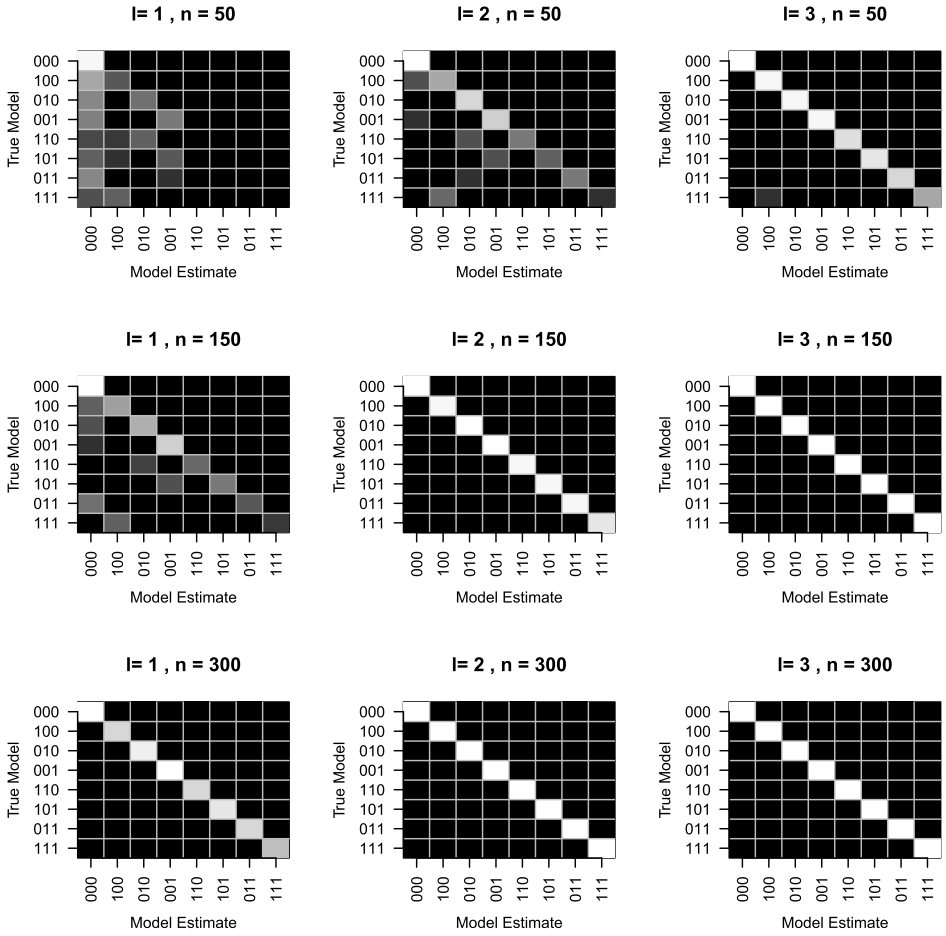
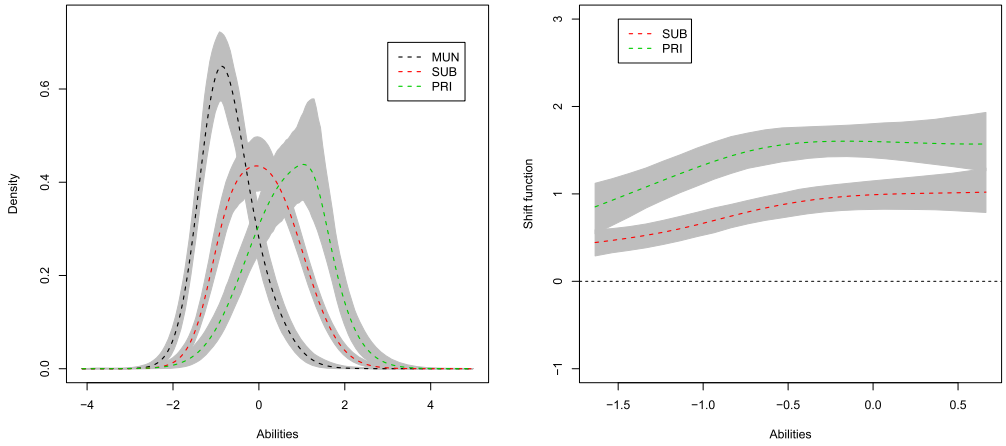


Figure 5: Posterior probabilities of the Monte Carlo simulation study of Section 4.2. Probability 0 is represented by black, while 1 is represented by white in the grayscale. The correctly identified hypotheses are represented in the main diagonal.

5 Applications

5.1 Educational achievement by school type

We implement our testing procedure to assess whether student educational achievement differs depending on the type of school they attend. In this context, it has been observed that, in some countries, school-type is a good proxy of students' socioeconomic-status. It is widely evidenced by international studies that educational achievement is strongly correlated with the socio-economic background of students (e.g., OECD, 2016). In Chile, this phenomenon has been present, not only in the case of international student assessments such as the Program for International Student Assessment (PISA), but also in



(a) (b)

Figure 6: (a) Posterior mean and 95% point-wise credibility intervals of the densities for the educational achievement by school type of Section 5.1. (b) Shift functions for the comparison of subsidized and private schools versus municipal (control).

national assessments such as SIMCE (Agencia de Calidad de la Educación, 2017) and SEPA (MIDE UC, 2017).

The data considered stems from the *System of Assessment Progress in Achievement* (SEPA, by its Spanish acronym); a private national evaluation system in Chile. SEPA consists of a battery of tests in the subjects of Mathematics and Language, designed to assess achievement in students from first to eleventh grade. In each application, besides students' test scores, additional information such as school type (e.g., municipal, subsidized, private) is also available in the data base. For the illustration, we consider the abilities on the mathematics test for a total of $n = 1,130$ students attending three different types of schools and distributed in the following way: $n_{Mun} = 324$, $n_{Sub} = 640$, $n_{Pri} = 166$; where n_{Mun}, n_{Sub}, and n_{Pri} are the sample sizes for the municipal, subsidized and private schools, respectively. The ability latent variables were predicted using a two parameter logistic item response theory model (2PL) (De Boeck and Wilson, 2004; van der Linden, 2016), with values defined in the real line.

Our aim here is to test whether ability distributions of the municipal schools (selected as the control group) are different to other types of schools. This selection is motivated by the interest of knowing how different are the achievements in private schools compared to the municipal ones, which constitute the public system. We fitted our model considering the same hyperparameters of Section 3.2 and 4.1. Figure 6 shows estimated ability distributions for each school type (dashed lines), 95% point-wise credibility intervals and the estimation of the shift functions that are used to compare the ability distributions of subsidized and private schools versus municipal ones. Visually, the ability distributions seem to differ from each other. Taking municipal schools as the control group, this

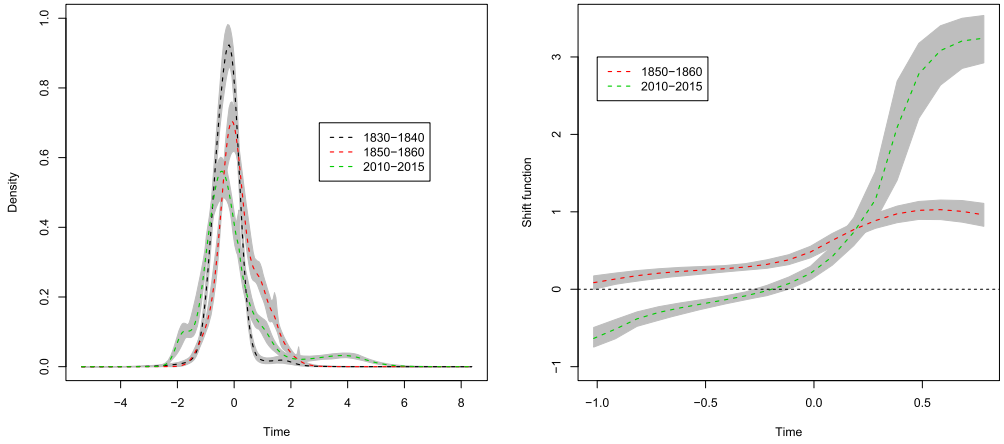
finding is formally confirmed by our model which gives support to the hypothesis H_{11} or, equivalently, assigned posterior probability concentrated in $\pi(H_{11} \mid \mathbf{y}) = 1$. The shifts functions show what was formally found using our BNP multiple testing procedure. As a matter of fact, none of the curves lie in the zero line, meaning that the ability distributions for schools differ. It can also be seen that the magnitude of the differences is different depending on the support region considered. For instance, if the students' abilities are low, there is some gain in changing type of school, and this gain is more evident if the abilities are higher. We also applied all the existing methods listed in Table 3. These methods found differences in locations or scales between the abilities distributions, but as can we seen in Figure 6 the differences are due to different kind of skewness. For instance, the municipal schools show positive skew, while the private schools show a negative skew distribution.

5.2 Phenology study

A consequence of climate change of increasing concern to ecologists is the decoupling of species interactions due to drastic changes in the timing of life cycle events. For example, large variations in the dates in which plants flower can have drastic impacts on species that depend upon them. The study of the timing of these events and how it is affected by variations in climate is called phenology.

Here we make use of historical data from a network of institutions collecting phenological data through participatory science methods between 1825–1878 in the state of New York, and compare it to a modern dataset between 2010 and 2015 in the same region. We restrict our analysis to a single phenological event across several species in the region, namely the *day of first flowering*. Taking the period between 1830–1840 as the control group ($n_c = 1,703$), we compare the first flowering dates between the control and two other time periods: 1850–1860 ($n_1 = 2,449$) and 2010–2015 ($n_2 = 2706$). The response variable is the (centered and scaled) day of the year, whose observed values were assumed to be iid conditionally on the time period. The goal is to determine if and how the distribution of the first flowering has varied through time. This was the reason that motivated the election of the first period of observation as the control group. However, we must disclose that this example is intended as a means to illustrate the capabilities of the proposed methods, rather than a rigorous attempt to determine the impact of climate change.

The results derived from our analysis indicate that the density of the first flowering during the control period markedly differs from the density of both the 1850–1860 decade and that from modern times, with the posterior probability entirely concentrating on this hypothesis $\pi(H_{11} \mid \mathbf{y}) = 1$. That being said, the regions of the support where these differences take place differ drastically between the two comparison groups (Figure 7). In particular, the results show that, when compared to the control group, the dates of first flowering occurred later during the year in the 1850–1860 period. In the 2010–2015 period more mass is assigned to both earlier and later flowering dates when compared to the control period, this behavior is clearly visualized from the shift function estimation. Notice that, the shift function allows us to identify which part of the population have a treatment effect. Conversely, as expected due to the large sample



(a) (b)

Figure 7: (a) Posterior mean and 95% point-wise credibility intervals for densities of the flowering dates of Section 5.2. Shift functions for the comparison of 1850–1860 decade and modern times versus 1830–1840 decade (control), (b).

size in this illustration, the existing methods considered in the Monte Carlo simulation section detected differences in location or scale, but in this study, the differences are not only due to location or scale, they are mainly given in the tails of the densities. Additionally, the existing methods do not provide a smooth estimation of the densities. Of course, simpler ways to estimate the density such as kernel estimators are available, but these kind of estimators are strongly affected by the selection of the bandwidth. On the other hand, the quantification of the uncertainty is not straightforward. In our proposal, the credible sets for the densities and shift function quantifies the uncertainty and allows us to visualize the differences.

6 Concluding remarks

We proposed a formal Bayesian hypothesis procedure to compare multiple treatments against a control. The methods developed are applicable to a wide variety of problems, and improve upon existing methods that test against a control group. The procedure avoids strong modeling assumptions over the distributions of each population, and is able to identify differences with respect to the entire distribution of the control. Additionally, we provide a simple approach to visualize the differences detected by the procedure between pairs of distributions by using the shift function.

The proposed method accounts for the multiplicity of the testing problem through the priors on the space of hypotheses. The comparisons are relatively simple, due to the nesting structure of the proposed model (see, Remark 1). The nesting structure is facilitated by considering common weights in the mixture model. Extensions of the

model to consider dependent weights are possible, given that a nested structure for the weights can be also specified. More flexibility could be added to the model considering for example a skew normal distribution for the kernel in (3), see, e.g. Canale and Scarpa (2016). An extension for multivariate responses could also be feasible if an adequate parametrization can be found for a multivariate normal kernel.

As shown in Section 4, the performance of our approach proved to be consistently good, in most cases outperforming the other alternatives considered in the Monte Carlo simulation study. Unsurprisingly, classical multiple testing procedures were only able to detect differences in locations. Similarly, the two-sample testing procedures exclusively detected the features for which they were designed. That is, t-test and Wilcoxon excelled at detecting differences in location; Levene successfully detected changes in scale; and Kolmogorov-Smirnov identified differences across the entire distribution. The conclusions derived from the Kolmogorov-Smirnov test are similar to those resulting from our method; however, our method has the advantage of being a multiple comparison procedure that yields density estimates for all populations, from which the shift function can be calculated. Of course, the proposed approach is not designed as a procedure to deal with hundreds or thousands of comparisons given the computational costs associated to it. The proposed strategy relaxes parametric assumptions, provides estimates for the strength of the competing hypotheses in the form of posterior probabilities, and has the potential to yield new insight through the use of the shift function.

For problems with small sample sizes, specific location or scale tests are preferred, given that they target specific features of the populations. Nevertheless, many current applications have sufficiently large data to make implementing our approach possible. Furthermore, Bayesian hypothesis testing procedures yield the wealth of information contained in the posterior probabilities, which can be combined with a loss function to make decisions as with classical tests.

Supplementary Material

Supplementary Material for 'A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control' (DOI: 10.1214/18-BA1122SUPP; .pdf). The online Supplementary Material contains the Gibbs Algorithm described in Section 3.4, as well as the image plots of the comparison between our proposal and other classical hypothesis tests (Section 4.2), including both multiple and two-sample cases.

References

Agencia de Calidad de la Educación (2017). "SIMCE." http://www.agenciaeducacion.cl/evaluaciones/que-es-el-simce/. 667

Altomare, D., Consonni, G., and La Rocca, L. (2013). "Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors." *Biometrics*, 69(2): 478–87. MR3071066. doi: https://doi.org/10.1111/biom.12018. 653

Barrientos, A. F., Jara, A., and Quintana, F. A. (2012). "On the support of MacEachern's dependent Dirichlet processes and extensions." *Bayesian Analysis*, 7: 277–310. MR2934952. doi: https://doi.org/10.1214/12-BA709. 655

Basu, S. and Chib, S. (2003). "Marginal likelihood and Bayes factors for Dirichlet Process Mixture models." *Journal of the American Statistical Association*, 98: 224–235. MR1965688. doi: https://doi.org/10.1198/01621450338861947. 653

Berger, J. and Pericchi, L. (1996). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91(433): 109–122. MR1394065. doi: https://doi.org/10.2307/2291387. 651, 653

Bhattacharya, A. and Dunson, D. (2012). "Nonparametric Bayes classification and hypothesis testing on manifolds." *Journal of Multivariate Analysis*, 111: 1–19. MR2944402. doi: https://doi.org/10.1016/j.jmva.2012.02.020. 652

Canale, A. and Scarpa, B. (2016). "Bayesian nonparametric location-scale-shape mixtures." *Test*, 25: 113–130. MR3463805. doi: https://doi.org/10.1007/s11749-015-0446-2. 670

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 654

Chen, Y. and Hanson, T. E. (2014). "Bayesian nonparametric k–sample tests for censored and uncensored data." *Computational Statistics and Data Analysis*, 71: 335–346. MR3131974. doi: https://doi.org/10.1016/j.csda.2012.11.003. 652

Cipolli, W., Hanson, T., and McLain, A. (2016). "Bayesian nonparametric multiple testing." *Computational Statistics and Data Analysis*, 101: 64–79. MR3504836. doi: https://doi.org/10.1016/j.csda.2016.02.016. 652

De Boeck, P. and Wilson, M. (2004). *Explanatory item response models. A generalized linear and nonlinear approach*. New York, USA: Springer. MR2083195. doi: https://doi.org/10.1007/978-1-4757-3990-9_2. 667

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99: 205–215. MR2054299. doi: https://doi.org/10.1198/016214504000000205. 652

Doksum, K. (1974). "Empirical probability plots and statistical inference for nonlinear models in the two-sample case." *The Annals of Statistics*, 2: 267–277. MR0356350. 652, 659

Doksum, K. and Sievers, G. (1976). "Plotting with confidence: Graphical comparisons of two populations." *Biometrika*, 63: 421–434. MR0443210. doi: https://doi.org/10.1093/biomet/63.3.421. 652

Dunnett, C. W. (1955). "A multiple comparison procedure for comparing several treatments with a control." *Journal of the American Statistical Association*, 50(272): 1096–1121. 650, 651

Dunnett, C. W. and Tamhane, A. C. (1991). "Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts." *Statistics in Medicine*, 10(6): 939–947. 650, 651, 664

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. MR1340510. 660

Gao, X., Alvo, M., Chen, J., and Li, G. (2008). "Nonparametric multiple comparison procedures for unbalanced one-way factorial designs." *Journal of Statistical Planning and Inference*, 138(8): 2574–2591. MR2432382. doi: https://doi.org/10.1016/j.jspi.2007.10.015. 651, 664

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 651, 653, 656

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. New York, USA: Springer. MR1992245. 651

Gutiérrez, L., Barrientos, A. F., González, J., and Taylor-Rodríguez, D. (2018). "Supplementary Material for 'A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control'." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1122SUPP. 658

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. (2010). *Bayesian nonparametrics*. Cambridge, UK: Cambridge University Press. MR2722988. doi: https://doi.org/10.1017/CBO9780511802478.002. 651

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. Wiley, New York. MR0914493. doi: https://doi.org/10.1002/9780470316672. 649

Hollander, M. and Korwar, R. (1980). *Nonparametric Bayesian Estimation of the Horizontal Distance Between Two Populations*. Defense Technical Information Center. 652

Hollander, M. and Wolfe, D. (1999). *Nonparametric Statistical Methods*. Wiley Series in Probability and Statistics. MR1666064. 651, 664

Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, 6(2): 65–70. MR0538597. 651

Holmes, C., Caron, F., Griffin, J., and Stephens, D. (2015). "Two–sample Bayesian nonparametric hypothesis testing." *Bayesian Analysis*, 10(2): 297–320. MR3420884. doi: https://doi.org/10.1214/14-BA914. 652

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: Frequentist and Bayesian strategies." *The Annals of Statistics*, 33(2): 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 653, 656

Jeffreys, H. (1935). "Some Tests of Significance, Treated by the Theory of Probability." *Proceedings of the Cambridge Philosophy Society*, 31: 203–222. 652

Jeffreys, H. (1961). *The theory of probability (3rd. Ed.)*. Oxford, UK: Oxford University Press. MR0187257. 651, 652, 654

Jeffreys, W. H. and Berger, J. O. (1992). "Ockham's razor and Bayesian analysis." *American Scientist*, 80(1): 64–72. 653

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2): 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 653

Johnson, V. E. and Rossell, D. (2012). "Bayesian Model Selection in High-Dimensional Settings." *Journal of the American Statistical Association*, 107(498): 649–660. MR2980074. doi: https://doi.org/10.1080/01621459.2012.682536. 653

Kalli, M., Griffin, J. E., and Walker, S. (2011). "Slice sampling mixture models." *Statistics and Computing*, 21: 93–105. MR2746606. doi: https://doi.org/10.1007/s11222-009-9150-y. 657

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90: 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 652

Kim, S., Dahl, D. B. D., and Vannucci, M. (2009). "Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models." *Bayesian Analysis*, 4(4): 707–732. MR2570085. doi: https://doi.org/10.1214/09-BA426. 652

Labadi, L. A., Masuadi, E., and Zarepour, M. (2014). "Two-sample Bayesian nonparametric goodness-of-fit test." *arXiv preprint arXiv:1411.3427*. 652

Ley, E. and Steel, M. F. (2009). "On the effect of prior assumptions in Bayesian model averaging with applications to growth regression." *Journal of applied econometrics*, 24(4): 651–674. MR2675199. doi: https://doi.org/10.1002/jae.1057. 654

Liang, F., Paulo, R., Molina, G., Clyde, M. a., and Berger, J. O. (2008). "Mixtures of g Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 651, 653

Lu, H. H. S., Wells, M. T., and Tiwari, R. C. (1994). "Inference for Shift Functions in the Two-Sample Problem With Right-Censored Data: With Applications." *Journal of the American Statistical Association*, 89(427): 1017–1026. MR1294747. 652

Ma, L. and Wong, W. H. (2011). "Coupling optional Pólya trees and the two sample problem." *Journal of the American Statistical Association*, 106: 1553–1565. MR2896856. doi: https://doi.org/10.1198/jasa.2011.tm10003. 652

MacEachern, S. N. (1999). "Dependent nonparametric processes." In *ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA*. American Statistical Association. 652, 655

MacEachern, S. N. (2000). "Dependent Dirichlet processes." Technical report, Department of Statistics, The Ohio State University. 652, 655

MIDE UC (2017). "SEPA." http://www.sepauc.cl/. 667

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 653

Müller, P. and Mitra, R. (2013). "Bayesian Nonparametric Inference–Why and How." *Bayesian Analysis*, 8(2): 269–302. MR3066939. doi: https://doi.org/10.1214/13-BA811. 651

Müller, P. and Quintana, F. (2004). "Nonparametric Bayesian data analysis." *Statistical Science*, 19: 95–110. MR2082149. doi: https://doi.org/10.1214/088342304000000017. 651

Müller, P., Quintana, F. A., and Rosner, G. (2004). "A method for combining inference across related nonparametric Bayesian models." *Journal of the Royal Statistical Society, Series B*, 66: 735–749. MR2088779. doi: https://doi.org/10.1111/j.1467-9868.2004.05564.x. 655

OECD (2016). "PISA 2015 Results (Volume I)." URL http://www.oecd.org/education/pisa-2015-results-volume-i-9789264266490-en.htm 666

Pati, D., Dunson, D. B., and Tokdar, S. T. (2013). "Posterior consistency in conditional distribution estimation." *Journal of Multivariate Analysis*, 116: 456–472. MR3049916. doi: https://doi.org/10.1016/j.jmva.2013.01.011. 655

Ramanan, G. and Berry, D. (1998). "A Bayesian multiple comparisons using Dirichlet process priors." *Journal of the American Statistical Association*, 93(443): 1130–1139. MR1649207. doi: https://doi.org/10.2307/2669856. 652

Ročková, V. and George, E. I. (2014). "EMVS: The EM approach to Bayesian variable selection." *Journal of the American Statistical Association*, 109(506): 828–846. MR3223753. doi: https://doi.org/10.1080/01621459.2013.869223. 653

Ročková, V. and George, E. I. (2016). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association*. MR3803476. doi: https://doi.org/10.1080/01621459.2016.1260469. 653, 656

Scott, J. G. (2009). "Nonparametric Bayesian multiple testing for longitudinal performance stratification." *The Annals of Applied Statistics*, 3(4): 1655–1674. MR2752152. doi: https://doi.org/10.1214/09-AOAS252. 652

Scott, J. G. and Berger, J. O. (2006). "An Exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136(7): 2144–2162. MR2235051. doi: https://doi.org/10.1016/j.jspi.2005.08.031. 652

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 651, 652, 653, 654

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 2: 639–650. MR1309433. 655

Simovici, D. A. and Djeraba, C. (2008). *Partially Ordered Sets*, 129–172. London: Springer London. 650

Soriano, J. and Ma, L. (2017). "Probabilistic multi–resolution scanning for two–sample differences." *Journal of the Royal Statistical Society: Series B*, 79: 547–572. MR3611759. doi: https://doi.org/10.1111/rssb.12180. 652

van der Linden, W. J. (ed.) (2016). *Handbook of item response theory. Three volume set*. Boca Raton, FL: Chapman and Hall/CRC. 667

Walker, S. (2007). "Sampling the Dirichlet mixture model with Slices." *Communications in Statistics- Simulation and Computation*, 36: 45–54. MR2370888. doi: https://doi.org/10.1080/03610910601096262. 657

Wells, M. T. and Tiwari, R. C. (1989). "Bayesian quantile plots and statistical inference for nonlinear models in the two sample case with incomplete data." *Communications in Statistics – Theory and Methods*, 18(8): 2955–2964. MR1033142. doi: https://doi.org/10.1080/03610928908830070. 652

Wilson, M., Iversen, E., Clyde, M. A., Schmidler, S. C., and Schildkraut, J. M. (2010). "Bayesian model search and multilevel inference for SNP association studies." *The Annals of Applied Statistics*, 4(3): 1342–1364. MR2758331. doi: https://doi.org/10.1214/09-AOAS322. 651, 654

Womack, A. J., Fuentes, C., and Taylor-Rodriguez, D. (2015). "Model Space Priors for Objective Sparse Bayesian Regression." *arXiv preprint arXiv:1511.04745*. 651, 654, 657

Womack, A. J., León-Novelo, L., and Casella, G. (2014). "Inference From Intrinsic Bayes' Procedures Under Model Selection and Uncertainty." *Journal of the American Statistical Association*, 109(507): 1040–1053. MR3265679. doi: https://doi.org/10.1080/01621459.2014.880348. 653

Zellner, A. and Siow, A. (1980). "Posterior odds ratios for selected regression hypotheses." In *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, 585–603. URL http://www.springerlink.com/index/5300770UP12246M9.pdf MR1483352. doi: https://doi.org/10.1214/lnms/1215452225. 653

Acknowledgments

The work of the first author was supported by grants Fondecyt 11140013, 'Proyecto REDES ETAPA INICIAL, Convocatoria 2017 REDI170094' and Iniciativa Científica Milenio – Minecon Núcleo Milenio MiDaS. The second author was supported by grants from the National Science Foundation (ACI 1443014 and SES 1131897) and the Alfred P. Sloan Foundation (G-2-15-20166003). The third author acknowledges partial support of the grant Fondecyt 1150233 and 'Proyecto REDES ETAPA INICIAL, Convocatoria 2017 REDI170094'. The authors thank the comments from the two referees and the associate editor, which contributed to improving the quality of the manuscript.