Finite-dimensional Discrete Random Structures and Bayesian Clustering

Antonio Lijoi[1,2], Igor Prünster[1,2] and Tommaso Rigon[3]

[1] Department of Decision Sciences, Bocconi University, Milano
[2] Bocconi Institute for Data Science and Analytics (BIDSA)
[3] Department of Economics, Management and Statistics, University of Milano-Bicocca

Abstract

Discrete random probability measures stand out as effective tools for Bayesian clustering. The investigation in the area has been very lively, with a strong emphasis on nonparametric procedures based on either the Dirichlet process or on more flexible generalizations, such as the normalized random measures with independent increments (NRMI). The literature on finite-dimensional discrete priors is much more limited and mostly confined to the standard Dirichlet-multinomial model. While such a specification may be attractive due to conjugacy, it suffers from considerable limitations when it comes to addressing clustering problems. In order to overcome these, we introduce a novel class of priors that arise as the hierarchical compositions of finite-dimensional random discrete structures. Despite the analytical hurdles such a construction entails, we are able to characterize the induced random partition and determine explicit expressions of the associated urn scheme and of the posterior distribution. A detailed comparison with (infinite-dimensional) NRMIs is also provided: indeed, informative bounds for the discrepancy between the partition laws are obtained. Finally, the performance of our proposal over existing methods is assessed on a real application where we study a publicly available dataset from the Italian education system comprising the scores of a mandatory nationwide test.

Keywords: Bayesian nonparametrics; Normalized random measures; Random partitions; Urn schemes.

1 Introduction

The definition and the investigation of flexible discrete priors, beyond the Dirichlet process, for Bayesian analysis has attracted increasing interest in recent years. Among the proposals that have been made, we recall the Pitman–Yor process (Pitman and Yor, 1997), the normalized inverse-Gaussian process (Lijoi et al., 2005), the normalized generalized gamma process (Lijoi et al., 2007) and the general classes of Gibbs-type priors (Gnedin and Pitman, 2005) and of normalized random measures with independent increments (NRMIs) (Regazzini et al., 2003). These priors have been mainly employed to address predictive inference and density estimation. More recently they have emerged as powerful tools to achieve sparsity in network models as effectively described in Caron and Fox (2017). While these are all instances of infinite-dimensional (nonparametric) priors, in this paper we embrace a different perspective and rely on a class of finite-dimensional priors whose flexibility can be increased at will, such that they can be seen as approximations of suitable nonparametric priors. Although this argument has been provably useful in Bayesian analysis (Green and Richardson, 2001; Miller and Harrison, 2018), it has mostly led to the specification of Dirichlet-based priors thus introducing in the analysis some limitations that are widely known in the literature. For example, the underlying model for the clustering structure is somehow restrictive, since it depends on a single parameter, the so-called concentration or total mass, thus calling for more flexible specifications. Such a narrow flexibility has further relevant effects on the associated system of predictive distributions. In the nonparametric case, these and other aspects are illustrated in De Blasi et al. (2015). Of course, one might mitigate this issue by placing a prior distribution on the concentration parameter, however the resulting posterior would not be available in closed form. Hence, albeit simple, such a specification has mostly turned out to be a sort of "black-box" prior. Besides these drawbacks, the lack of theoretical results has also prevented the development of computational algorithms that may facilitate their use. For example, while our investigation leads to exact sampling from the posterior distribution, a prior on the concentration parameter in a Dirichlet-based structure would require a Metropolis step in the computation.

The main focus of our paper will be clustering and density estimation. Just to set up some preliminary notation, let $(X_n)_{n\geq 1}$ be a sequence of $\mathbb{X}$-valued exchangeable random elements and let $K : \mathbb{X} \times \Theta \to \mathbb{R}^+$ be a transition kernel such that $x \mapsto K(x;\theta)$ is a density

function on $\mathbb{X}$, for any $\theta \in \Theta$. Conditionally on a random probability measure $\tilde{p}$, we suppose that $(X_i \mid \tilde{p}) \overset{\text{iid}}{\sim} \int_\Theta K(x;\theta)\tilde{p}(\mathrm{d}\theta)$ and

$$\tilde{p} = \sum_{h=1}^{H} \pi_h \delta_{\tilde{\theta}_h}, \qquad H \in \{1, 2, \ldots, \infty\}, \tag{1}$$

where $\delta_{\tilde{\theta}_h}$ denotes the unit mass at $\tilde{\theta}_h$, $\sum_{h=1}^{H} \pi_h = 1$ almost surely and the collections $\{\tilde{\theta}_1, \ldots, \tilde{\theta}_H\}$ and $\{\pi_1, \ldots, \pi_H\}$ are independent. Furthermore, $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$ are iid draws from a *diffuse* probability measure P. Hence, $\tilde{p}$ is a proper species sampling model (Pitman, 1996). It is worth noting that the theoretical results that are displayed in the next sections are essential for applying our proposal to a number of other inferential problems, beyond the specific ones we consider here. With $H < \infty$, the symmetric DIRICHLET$(c/H, \ldots, c/H)$ distribution is a popular prior choice for $(\pi_1, \ldots, \pi_H)$ and it is well-known that, in this case, the law of $\tilde{p}$ in (1) weakly converges to the law of the Dirichlet process, with parameters (c, P), as $H \to \infty$. Such a symmetric Dirichlet specification, when directly used for exchangeable data $(X_1, \ldots, X_n \mid \tilde{p}) \overset{\text{iid}}{\sim} \tilde{p}$, is often referred to as *Dirichlet multinomial model*. See for instance Ishwaran and Zarepour (2002). It is worth recalling that this model may be used for sparse finite mixtures as a way to circumvent the issue of selecting the number of mixture components (Malsiner-Walli et al., 2016). Indeed, under suitable assumptions, overfitted mixtures consistently select the correct number of mixture components (Rousseau and Mengersen, 2011). See also Nguyen (2013) for further asymptotic validations of mixture models based on Dirichlet-like mixing measures. In addition, the Dirichlet multinomial process has been employed in a variety of statistical applications, e.g. for Bayesian modeling of network data (Durante et al., 2017; Durante and Dunson, 2018), for semi-parametric random effects in regression models (Rigon et al., 2019), and for functional data analysis (Dunson et al., 2008; Petrone et al., 2009). Finally, the symmetric Dirichlet for $(\pi_1, \ldots, \pi_H)$ may be combined with a "repulsive" prior for the locations $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$ as in Xu et al. (2016).

In this paper we introduce and study a general class of finite-dimensional random probability measures $\tilde{p}$ as in (1), which will be termed *normalized infinitely divisible multinomial processes* (NIDM). The investigation we develop will benefit from the connection with homogeneous NRMIs, whose theory has achieved remarkable advances in the recent years (Lijoi and Prünster, 2010). The proposed NIDM processes allow for a finer control of the under-

lying clustering mechanism and for a robustification of the prior specification process. We will provide both theoretical and empirical evidence in support of this claim, in line with what was already noticed in the infinite-dimensional case (Lijoi et al., 2007). As a by-product of our investigation, we note that one might employ NIDMs as approximations of their infinite-dimensional counterpart. Besides the theoretical interest that a result of this type may give rise to, it is also very helpful from a practical standpoint since it helps lightening computational bottlenecks. Indeed, posterior inference for NRMIs most often relies on series representations of the type devised in Ferguson and Klass (1972), hence requiring numerical and analytical approximations. In contrast, the posterior structure of several NIDMs can be computed and sampled exactly, without the need of MCMC steps. This is an important bonus, even when compared to standard Dirichlet-multinomial specifications. Such a gain, however, does not come for free since the probabilistic structure of our model yields some challenging technical hurdles when it comes to determining distributional properties of interest for Bayesian inference. These difficulties parallel those that arise when one uses a discrete base measure for a nonparametric prior process. See, e.g., Canale et al. (2017) for an example in the Pitman–Yor case.

The main application we focus on concerns the INVALSI 2016-2017 dataset, a national examination conducted in Italy. Specifically, we aim at measuring the teaching competencies of a set of schools by taking into account the socio-demographic characteristics of its students. Great effort has been made by the public research institute INVALSI to provide reliable quantifications of the effect of each school on the test performance. Indeed, such an indicator is nationally relevant especially for the development and the evaluation of educational policies. We address this problem via semi-parametric modeling with nonparametric school-specific random effects, which will be interpreted as a proxy of the added-value of the school.

The paper is organized as follows. In Section 2 we review some background material about completely random measures and species sampling models. In Section 3 we define NIDM processes and we discuss several *a priori* properties, such as the law of the random partition they induce, and the weak convergence to NRMIs. In Section 4, we discuss generalized urn schemes and posterior characterizations, with a particular emphasis on the normalized generalized gamma (NGG) multinomial process. In Section 5 we employ the NGG multinomial prior for the analysis of a real dataset, highlighting practical advantages over existing methods. We

conclude with a final discussion in Section 6. Detailed proofs are deferred to the supplementary material. This further includes an extensive simulation study that shows the greater flexibility of the NGG multinomial prior under a wide range of data generating distributions, with different choices of the hyperparameters and sample sizes.

2 Discrete priors and random partitions

Since NIDM processes are closely related to homogeneous NRMIs, we sketch a concise overview of some preliminary definitions and notation on random measures.

Let Θ be a separable and complete metric space and let $\mathscr{B}(\Theta)$ be its Borel σ-field. A random measure $\tilde{\mu}_\infty$ on $(\Theta, \mathscr{B}(\Theta))$ such that $\tilde{\mu}_\infty(A_1), \ldots, \tilde{\mu}_\infty(A_d)$ are mutually independent random variables for any choice of pairwise disjoint $A_1, \ldots, A_d$ in $\mathscr{B}(\Theta)$, and for any $d \geq 2$, is a *completely random measure* (CRM). For our purposes, it is useful to recall that a CRM with no fixed points of discontinuity and no deterministic drift can be represented as $\tilde{\mu}_\infty = \sum_{h=1}^{\infty} J_h \delta_{\tilde{\theta}_h}$ and is characterized by the Laplace functional transform

$$\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_\infty(\mathrm{d}\theta)}\right) = \exp\left\{-\int_{\mathbb{R}^+\times\Theta}\left(1-\mathrm{e}^{-sf(\theta)}\right)\nu(\mathrm{d}s,\mathrm{d}\theta)\right\}, \tag{2}$$

where $f : \Theta \to \mathbb{R}^+$ is a measurable function and the measure ν on $\mathbb{R}^+ \times \Theta$, termed Lévy measure, or intensity, characterizes the CRM and is such that $\int_{\mathbb{R}^+\times A} \min\{1, s\}\nu(\mathrm{d}s, \mathrm{d}\theta) < \infty$ for any bounded $A \in \mathscr{B}(\Theta)$. In the following, we consider only *homogeneous* CRMs, which amounts to having a Lévy intensity of the form $\nu(\mathrm{d}s, \mathrm{d}\theta) = \rho(s)\mathrm{d}s\, cP(\mathrm{d}\theta)$, where P is a probability measure over $(\Theta, \mathscr{B}(\Theta))$ and $c \in \mathbb{R}^+$. We will use the notation $\tilde{\mu}_\infty \sim \mathrm{CRM}(c, \rho; P)$. If one additionally has that $\int_{\mathbb{R}^+} \rho(s)\,\mathrm{d}s = \infty$, then $0 < \tilde{\mu}(\Theta) < \infty$ almost surely, and a homogeneous NRMI is defined as $\tilde{p}_\infty = \sum_{h=1}^{\infty}(J_h/\bar{J})\delta_{\tilde{\theta}_h}$, where $\bar{J} = \sum_{h=1}^{\infty} J_h = \tilde{\mu}(\Theta)$. We will write $\tilde{p}_\infty \sim \mathrm{NRMI}(c, \rho; P)$ to denote such a random probability measure. Several relevant nonparametric priors are NRMIs and a noteworthy class, which is the object of investigation in the present paper, arises when

$$\rho(s) = \frac{1}{\Gamma(1-\sigma)}\, s^{-1-\sigma}\, \mathrm{e}^{-\kappa s}, \tag{3}$$

whose additional parameters $0 \leq \sigma < 1$ and $\kappa \geq 0$ are such that at least one of them

is positive (Brix, 1999). The resulting NRMI is often referred to as *normalized generalized gamma* (NGG) process, and it includes some well-known nonparametric priors as special cases. See Lijoi et al. (2007).

Both NRMIs and the novel class of NIDM processes are discrete random probability measures. Thus, when their law identifies the de Finetti measure of the exchangeable sequence of Θ-valued random elements $(\theta_n)_{n\geq 1}$, there will be ties with positive probability, namely $\mathbb{P}[\theta_i = \theta_{i'}] > 0$ for any $i \neq i'$. Hence, an n-sample $\boldsymbol{\theta}^{(n)} = (\theta_1, \ldots, \theta_n)$ will display $K_n = k \leq n$ distinct values, say $\theta_1^*, \ldots, \theta_k^*$, with respective frequencies $n_1, \ldots, n_k$, so that $\sum_{j=1}^k n_j = n$. This amount to saying that $\boldsymbol{\theta}^{(n)}$ induces a random partition Ψ_n of $[n] = \{1, \ldots, n\}$ into k sets $C_1, \ldots, C_k$ such that $i \in C_j$ if and only if $\theta_i = \theta_j^*$. As discussed in Pitman (1996), this clustering mechanism is regulated by a symmetric function called *exchangeable partition probability function* (EPPF), which is defined by

$$\mathbb{P}(\Psi_n = \{C_1, \ldots, C_k\}) = \Pi(n_1, \ldots, n_k) = \sum_{i_1 \neq \ldots \neq i_k} \mathbb{E}\left(\prod_{j=1}^k \pi_{i_j}^{n_j}\right), \tag{4}$$

where the vector $\boldsymbol{n}_k = (n_1, \ldots, n_k)$ of positive integers is such that $n_j = \text{card}(C_j)$ and $\sum_{j=1}^k n_j = n$. Notice further that the conditional probabilities that θ_{n+1} displays a new value from the diffuse base measure P or coincides with a previously observed θ_j^*, given $\boldsymbol{\theta}^{(n)}$, can be expressed in terms of the EPPF as follows $w_0(\boldsymbol{n}^{(k)}) = \Pi(n_1, \ldots, n_k, 1)/\Pi(n_1, \ldots, n_k)$ and $w_j(\boldsymbol{n}^{(k)}) = \Pi(n_1, \ldots, n_j + 1, \ldots, n_k)/\Pi(n_1, \ldots, n_k)$, for $j = 1, \ldots, k$, which, for any $A \in \mathscr{B}(\Theta)$, entails $\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}^{(n)}) = w_0(\boldsymbol{n}^{(k)})P(A) + \sum_{j=1}^k w_j(\boldsymbol{n}^{(k)})\delta_{\theta_j^*}(A)$. If $\tilde{p}_\infty \sim \text{NRMI}(c, \rho; P)$, with P diffuse, the corresponding EPPF is

$$\Pi_\infty(n_1, \ldots, n_k) = \frac{c^k}{\Gamma(n)} \int_{\mathbb{R}^+} u^{n-1}\, \mathrm{e}^{-c\psi(u)} \prod_{j=1}^k \tau_{n_j}(u)\, \mathrm{d}u, \tag{5}$$

where $\psi(u) = \int_{\mathbb{R}^+}(1 - \mathrm{e}^{-us})\rho(s)\mathrm{d}s$ is the *Laplace exponent* and $\tau_m(u) := \int_{\mathbb{R}^+} s^m \mathrm{e}^{-us}\rho(s)\,\mathrm{d}s$, for any integer $m \geq 1$. If ρ is as in (3), i.e. $\tilde{p}_\infty$ is a NGG process, one finds out that $\Pi_\infty(n_1, \ldots, n_k) = \mathscr{V}_{n,k} \prod_{j=1}^k (1 - \sigma)_{n_j - 1}$, with $(a)_n = a(a+1)\cdots(a+n-1)$, for any real a

and integer $n \geq 1$, $(a)_0 = 1$, and

$$\mathscr{V}_{n,k} := \frac{\mathrm{e}^{\beta}\sigma^{k-1}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}; \beta\right), \tag{6}$$

where $\beta = c\kappa^{\sigma}/\sigma$, and $\Gamma(x; a) = \int_x^{\infty} s^{a-1}\mathrm{e}^{-s}\mathrm{d}s$ for any $a > 0$ is the incomplete gamma function. One may refer to Lijoi et al. (2007) for details about the determination of (6).

3 Normalized infinitely divisible multinomial processes

In order to define a more flexible, and still tractable, class of finite-dimensional priors we crucially start from infinitely divisible masses and adopt a normalization procedure similar to the one that yields NRMIs.

3.1 NIDM processes

The main building block of the new class of processes that we are proposing is a collection of independent and infinitely divisible random variables. In particular, we will deal with *finite* and *strictly positive* infinitely divisible random variables, say J, such that

$$\mathbb{E}\left(\mathrm{e}^{-\lambda J}\right) = \exp\{-c\psi(\lambda)\} = \exp\left\{-c\int_{\mathbb{R}^+} (1 - \mathrm{e}^{-\lambda s})\rho(s)\,\mathrm{d}s\right\}, \tag{7}$$

for any $\lambda > 0$ and some constant $c > 0$. The function ρ is a non-negative measurable function satisfying the same conditions outlined in Section 2, i.e. $\int_{\mathbb{R}^+} \min\{1, s\}\rho(s)\,\mathrm{d}s < \infty$ and $\int_{\mathbb{R}^+} \rho(s)\,\mathrm{d}s = \infty$. We will henceforth use the notation $J \sim \mathrm{ID}(c, \rho)$.

Definition 1. A random measure $\tilde{\mu}_H$ on $(\Theta, \mathscr{B}(\Theta))$ such that

$$\tilde{\mu}_H = \sum_{h=1}^{H} J_h \delta_{\tilde{\theta}_h}, \qquad J_h \overset{\text{iid}}{\sim} \mathrm{ID}\left(\frac{c}{H}, \rho\right), \qquad \tilde{\theta}_h \overset{\text{iid}}{\sim} P, \tag{8}$$

where $H < \infty$ and P is a *diffuse* probability measure on $(\Theta, \mathscr{B}(\Theta))$, is a *multinomial* random measure with *infinitely divisible* jumps and will be denoted as $\tilde{\mu}_H \sim \mathrm{IDM}(c, \rho; P)$.

It is important to stress that IDM's are not completely random, since $\tilde{\mu}_H(A_1), \ldots, \tilde{\mu}_H(A_d)$ are not mutually independent random variables for any choice of pairwise disjoint sets $A_1, \ldots, A_d$ in $\mathscr{B}(\Theta)$ and for any $d \geq 2$. Indeed, if $A \in \mathscr{B}(\Theta)$ is such that $0 < P(A) < 1$, then $\mathbb{P}(\tilde{\mu}_H(A) = 0) = (1 - P(A))^H > 0$. On the other hand, $\mathbb{P}(\tilde{\mu}_H(A) = 0 \mid \tilde{\mu}_H(A^c) = 0) = 0$ since $\mathbb{P}(\tilde{\mu}_H(\Theta) > 0) = 1$. Hence, independence between $\tilde{\mu}_H(A)$ and $\tilde{\mu}_H(A^c)$ does not hold true. Nonetheless, the Laplace functional transform of a IDM is available and, for any non–negative function f, equals

$$\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_H(\mathrm{d}\theta)}\right) = \left(\int_\Theta \exp\left\{-\frac{c}{H}\int_{\mathbb{R}^+}\left(1-\mathrm{e}^{-sf(\theta)}\right)\rho(s)\mathrm{d}s\right\} P(\mathrm{d}\theta)\right)^H. \tag{9}$$

Now set $f(\theta) = \lambda I_A(\theta)$ with $\lambda > 0$, $A \in \mathscr{B}(\Theta)$ and I_A denoting the indicator function of A. Then $\mathbb{E}\left(\mathrm{e}^{-\lambda\tilde{\mu}_H(A)}\right) = \left(1 - P(A) + P(A)\exp\left\{-c\psi(\lambda)/H\right\}\right)^H$, where ψ is the Laplace exponent defined in (7). It is apparent that $\tilde{\mu}_H(A)$ equals in distribution a binomial compound random element, namely $(\tilde{\mu}_H(A) \,|\, \tilde{m}) \overset{\mathrm{d}}{=} \sum_{j=0}^{\tilde{m}} J_j$ with $\tilde{m} \sim$ BINOMIAL$(H, P(A))$ where, conditional on $\tilde{m}$, the random variables $J_1, \ldots, J_{\tilde{m}}$ are iid from ID$(c/H, \rho)$ and J_0 is a point mass at zero: this provides an interesting and alternative representation of $\tilde{\mu}_H(A)$.

The new class of finite-dimensional processes that we define are obtained by normalizing random measures in Definition 1. In this respect, the construction is related to the normalized infinitely divisible (NID) family of distributions in Favaro et al. (2011), in the sense that the random probability weights are obtained by normalizing infinitely divisible random variables.

Definition 2. Let $\tilde{\mu}_H \sim$ IDM$(c, \rho; P)$. If $\tilde{p}_H = \tilde{\mu}_H/\tilde{\mu}_H(\Theta) = \sum_{h=1}^H \pi_h \delta_{\tilde{\theta}_h}$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_H) = (J_1/\sum_{h=1}^H J_h, \ldots, J_H/\sum_{h=1}^H J_h)$, then $\tilde{p}_H$ is a *normalized infinitely divisible multinomial process* and is denoted as $\tilde{p}_H \sim$ NIDM$(c, \rho; P)$. Notice, further, that $\boldsymbol{\pi}$ is a *normalized infinitely divisible* (NID) random vector and we use the notation $(\pi_1, \ldots, \pi_{H-1}) \sim$ NID$(c/H, \ldots, c/H; \rho)$ to identify it.

Hence, a NIDM process is a random probability measure with finite support whose vector of probability weights $\boldsymbol{\pi}$ is termed NID according to the terminology in Favaro et al. (2011). Indeed, Favaro et al. (2011) propose this class of distributions on the simplex as an alternative to the Dirichlet distribution. While they investigate some distributional properties and discuss examples for which the density function of $\boldsymbol{\pi}$ is available in closed form, they neither consider their uses in a Bayesian setting nor construct random probability measures with fi-

nite support based on $\boldsymbol{\pi}$. From Definition 2, it is apparent that $\tilde{\mu}(\Theta) \sim \text{ID}(c, \rho)$, thus ensuring that the above normalization is well-defined. Note that a NIDM process is also a proper species sampling model (1) with $H < \infty$, since P is diffuse. If we set $\rho(s) = s^{-1}\mathrm{e}^{-s}$ we recover the Dirichlet multinomial process, whose weights $(\pi_1, \ldots, \pi_{H-1}) \sim \text{DIRICHLET}(c/H, \ldots, c/H)$. Henceforth, we take ρ as in (3) and obtain a novel and tractable class of NIDM processes that we will term NGG multinomial process.

It is useful to point out that NIDM processes display a hierarchical structure. Indeed, if $(\tilde{\mu}_H \mid \tilde{p}_{0,H}) \sim \text{CRM}(c, \rho; \tilde{p}_{0,H})$ and $\tilde{p}_{0,H} = (H)^{-1} \sum_{h=1}^{H} \delta_{\tilde{\theta}_h}$, with $\tilde{\theta}_h \overset{\text{iid}}{\sim} P$, then $\tilde{\mu}_H \sim \text{IDM}(c, \rho; P)$. Hence, if we pick $\tilde{p}_H = \tilde{\mu}_H / \tilde{\mu}_H(\Theta)$ one has that $\tilde{p}_H \sim \text{NIDM}(c, \rho; P)$ and in view of the previous remark, it can be described through the following hierarchical model

$$(\tilde{p}_H \mid \tilde{p}_{0,H}) \sim \text{NRMI}(c, \rho; \tilde{p}_{0,H}), \qquad \tilde{p}_{0,H} = \frac{1}{H} \sum_{h=1}^{H} \delta_{\tilde{\theta}_h}. \tag{10}$$

The converse holds true as well: any NIDM can be represented as a hierarchical process with a NRMI having a discrete baseline measure at the bottom of the hierarchy. Note that the law of $\tilde{p}_{0,H}$ is that of a specific Gibbs–type prior, arising when the discount parameter goes to $-\infty$; see Gnedin and Pitman (2005). Furthermore, this representation relates any NIDM process to hierarchical constructions like those presented in Camerlenghi et al. (2018).

In view of (10), it is instructive to compare the finite-dimensional distributions of a NIDM $\tilde{p}_H$ with those of a $\tilde{p}_\infty \sim \text{NRMI}(c, \rho; P)$. In particular, it follows that, for any finite partition $\{B_1, \ldots, B_d\}$ of Θ into $\mathscr{B}(\Theta)$-sets, the distribution of $(\tilde{p}_H(B_1), \ldots, \tilde{p}_H(B_{d-1}))$ can be expressed as a mixture of a NID distribution with multinomial weights, motivating the NIDM denomination. More precisely,

$$\begin{aligned}(\tilde{p}_H(B_1), \ldots, \tilde{p}_H(B_{d-1})) \mid (\tilde{m}_1, \ldots, \tilde{m}_d) &\sim \text{NID}\left(c\,\frac{\tilde{m}_1}{H}, \ldots, c\,\frac{\tilde{m}_d}{H};\ \rho\right) \\ (\tilde{m}_1, \ldots, \tilde{m}_d) &\sim \text{MULTINOMIAL}\left(H, (P(B_1), \ldots, P(B_d))\right).\end{aligned}$$

On the other hand, since $\tilde{p}_\infty(B_i) = J_i / \sum_{j=1}^{d} J_j$, where $J_i = \tilde{\mu}_\infty(B_i) \sim \text{ID}(cP(B_i); \rho)$, one has $(\tilde{p}_\infty(B_1), \ldots, \tilde{p}_\infty(B_{d-1})) \sim \text{NID}\left(cP(B_1), \ldots, cP(B_d); \rho\right)$. As $H \to \infty$, an application of the law of large numbers yields $(c\tilde{m}_1/H, \ldots, c\tilde{m}_d/H) \xrightarrow{\text{a.s.}} (cP(B_1), \ldots, cP(B_d))$. A similar argument will be used in the next section to show that NIDM processes weakly converge to

the corresponding homogeneous NRMI as $H \to \infty$. Finally, note that the moments of a NIDM process can be obtained in closed form; see the supplementary material for details.

3.2 Convergence of NIDM processes

The previous discussion suggests that, when H is large enough, a NIDM approaches a non-parametric prior that corresponds to a homogeneous NRMI. The main result of this Section concerns the convergence of IDM random measures to homogeneous CRMs and we henceforth resort to the more concise notation $\tilde{\mu}(f) = \int_\Theta f(\theta)\, \tilde{\mu}(\mathrm{d}\theta)$.

Theorem 1. *Let* $\tilde{\mu}_H \sim \text{IDM}(c, \rho; P)$ *and* $\tilde{\mu}_\infty \sim \text{CRM}(c, \rho; P)$. *Then as* $H \to \infty$, *one has* $\mathbb{E}\left(\mathrm{e}^{-\tilde{\mu}_H(f)}\right) \to \mathbb{E}\left(\mathrm{e}^{-\tilde{\mu}_\infty(f)}\right)$ *for any positive and measurable function* $f : \Theta \to \mathbb{R}^+$ *such that* $\int_\Theta \psi(f(\theta))P(\mathrm{d}\theta) < \infty$.

Note that if f in the above theorem is integrable with respect to P then the condition $\int_\Theta \psi(f(\theta))P(\mathrm{d}\theta) < \infty$ is satisfied. In addition, given the convergence of the Laplace functionals, vague convergence $\tilde{\mu}_H \xrightarrow{v} \tilde{\mu}_\infty$ as $H \to \infty$ is implied by Kallenberg (Theorem 4.11, 2017) since the above condition is satisfied if f is a positive, continuous and bounded function. Now recall that $\tilde{p}_H \sim \text{NIDM}(c, \rho; P)$ and that $\tilde{p}_\infty \sim \text{NRMI}(c, \rho; P)$. An important implication of Theorem 1 is the convergence of the finite-dimensional distributions of $\tilde{p}_H$ to those of $\tilde{p}_\infty$. Indeed, plugging in Theorem 1 the simple function $f(\theta) = \sum_{i=1}^d \lambda_i I_{A_i}(\theta)$, for any collection of sets $A_1, \ldots, A_d \in \mathscr{B}(\Theta)$ and positive constants $\lambda_1, \ldots, \lambda_d > 0$, one has that $(\tilde{p}_H(A_1), \ldots, \tilde{p}_H(A_d)) \xrightarrow{d} (\tilde{p}_\infty(A_1), \ldots, \tilde{p}_\infty(A_d))$, as $H \to \infty$, as a consequence of the continuous mapping theorem. When working with random probability measures, the convergence of the finite-dimensional distributions suffices to guarantee the weak convergence of the whole process, which will be indicated with the $\xrightarrow{w}$ notation; see Kallenberg (Theorem 4.11, 2017).

Corollary 1. *Let* $\tilde{p}_H \sim \text{NIDM}(c, \rho; P)$ *and* $\tilde{p}_\infty \sim \text{NRMI}(c, \rho; P)$. *Then* $\tilde{p}_H \xrightarrow{w} \tilde{p}_\infty$ *as* $H \to \infty$.

The above statement implies the convergence in distribution of general functionals $\tilde{p}_H(f) \xrightarrow{d} \tilde{p}_\infty(f)$ as $H \to \infty$ when f is a continuous and bounded function. In the Dirichlet multinomial process case, related results were previously obtained, e.g., in Kingman (1975) and Green and Richardson (2001).

3.3 Random partitions and number of clusters

In this section we study the clustering mechanism underlying a NIDM process that amounts to determining the EPPF, namely the probability distribution of the induced exchangeable partition: this will be denoted by Π_H and it is defined through (4), with $\tilde{p}$ being replaced by $\tilde{p}_H$. To be more specific, it will be shown that the EPPF of any NIDM process is a finite mixture of the partition functions arising in the infinite-dimensional setting. Before stating the theorem, let us introduce some further quantity of interest. Define for any $m \geq 1$

$$\mathcal{V}_{m,H}(u) := \left((-1)^m \frac{\partial^m}{\partial u^m} \mathrm{e}^{-\frac{c}{H}\psi(u)}\right) \mathrm{e}^{\frac{c}{H}\psi(u)} = \frac{c}{H}\Delta_{m,H}(u),$$

and set $\mathcal{V}_{0,H} := 1$, where $\psi(u)$ is the Laplace exponent defined as in (7) and

$$\Delta_{m,H}(u) = \sum_{\ell=1}^{m} \left(\frac{c}{H}\right)^{\ell-1} \frac{1}{\ell!} \sum_{\boldsymbol{q}} \binom{m}{q_1 \ \cdots \ q_\ell} \prod_{r=1}^{\ell} \int_0^\infty s^{q_r} \mathrm{e}^{-us} \rho(s) \, \mathrm{d}s,$$

with the inner sum running over all vectors $\boldsymbol{q} = (q_1, \ldots, q_\ell)$ of positive integers such that $\sum_{r=1}^{\ell} q_r = m$. Moreover, for any vector $\boldsymbol{x} \in \mathbb{R}^p$, we let $|\boldsymbol{x}| = \sum_{i=1}^{p} x_i$.

Theorem 2. *Let $(\theta_n)_{n\geq 1}$ be an exchangeable sequence directed by a* NIDM *process prior. Then, the associated* EPPF *when $k \leq \min\{n, H\}$ is given by*

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{H^k (H-k)!} \frac{c^k}{\Gamma(n)} \int_{\mathbb{R}^+} u^{n-1} \mathrm{e}^{-c\psi(u)} \prod_{j=1}^{k} \Delta_{n_j,H}(u) \mathrm{d}u,$$

Moreover, if Π_∞ is the EPPF *of the corresponding homogeneous* NRMI *in* (4)*, one has*

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{H^k (H-k)!} \sum_{\boldsymbol{\ell}} \frac{1}{H^{|\boldsymbol{\ell}|-k}} \prod_{j=1}^{k} \frac{1}{\ell_j!} \sum_{\boldsymbol{q}_j} \binom{n_j}{q_{j1}, \ldots, q_{j\ell_j}} \times \Pi_\infty(q_{11}, \ldots, q_{1\ell_1}, \ldots, q_{k1}, \ldots, q_{k\ell_k}), \quad (11)$$

where the first sum runs over all vectors $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ such that $\ell_j \in \{1, \ldots, n_j\}$, and the jth of the k sums runs over $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{j\ell_j})$ such that $q_{jr} \geq 1$ and with $|\boldsymbol{q}_j| = n_j$.

This mixture representation in (11) is reminiscent of the one in Camerlenghi et al. (2018) for hierarchical NRMIs, which is not surprising in view of the hierarchical representation of NIDM processes in (10). Furthermore, note that $\lim_{H\to\infty} \Delta_{m,H}(u) = \tau_m(u)$ for any $u > 0$ and $m \geq 1$. This leads to show that $\lim_{H\to\infty} \Pi_H = \Pi_\infty$, namely the EPPF associated to a NIDM process converges to the one induced by a homogeneous NRMI and displayed in (5). As a matter of fact, one can say something more precise and identify bounds for the ratio between Π_H and Π_∞, for any H, as stated in the next theorem.

Theorem 3. *For any $k \leq \min\{n, H\}$ one has*

$$\frac{H!}{H^k(H-k)!} \leq \frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_\infty(n_1,\ldots,n_k)} \leq \int_{\mathbb{R}^+} \prod_{j=1}^{k} \frac{\Delta_{n_j,H}(u)}{\tau_{n_j}(u)} q_\infty(u)\mathrm{d}u,$$

where $q_\infty(u) \propto u^{n-1}\mathrm{e}^{-c\psi(u)} \prod_{j=1}^{k} \tau_{n_j}(u)$ is a density function.

Note that q_∞ is the density function of a latent random variable that is used in James et al. (2009) to provide posterior characterizations of NRMIs. Both bounds converge to 1 as $H \to \infty$. As a simple application of Theorem 3, one might obtain bounds also for the predictive distributions, by exploiting their relationship with the EPPF. For NGG multinomial processes and, a fortiori, in the Dirichlet multinomial case, the EPPF and the related bounds can be computed explicitly. This is illustrated in the following examples.

Example 1 (Dirichlet multinomial process)**.** Let the NIDM process be characterized by the intensity function $\rho(s) = s^{-1}\mathrm{e}^{-s}$. For any $k \leq \min\{n, H\}$, in view of Theorem 2 one has $\Pi_H(n_1,\ldots,n_k) = H!/(H-k)!(c)_n^{-1} \prod_{j=1}^{k} (c/H)_{n_j}$. This EPPF can be found, e.g., in Green and Richardson (2001). Additionally, note that such a prior is a Pitman–Yor process with discount parameter $\sigma = -c/H < 0$ and $H|\sigma| = c > 0$. See, e.g., De Blasi et al. (2015). Straight application of Theorem 3 yields

$$\frac{H!}{H^k(H-k)!} \leq \frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_\infty(n_1,\ldots,n_k)} \leq \prod_{j=1}^{k} \frac{(1+c/H)_{n_j-1}}{(n_j-1)!}.$$

This makes clear that Π_H and Π_∞ are close when either H is large, as it is natural to expect on the basis of Corollary 1, or the total mass parameter c is small. Note that this is the same

bound obtained in the Appendix of Ishwaran and Zarepour (2002) by means of different techniques, and thus Theorem 3 can be seen as a generalization of their result.

Example 2 (NGG multinomial process)**.** Let the NIDM process be characterized by the generalized gamma intensity function $\rho(s)$ given in (3). On the basis of Theorem 2, for any $k \leq H$ one has

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!} \sum_{\boldsymbol{\ell}} \frac{\mathscr{V}_{n,|\boldsymbol{\ell}|}}{H^{|\boldsymbol{\ell}|}} \prod_{j=1}^{k} \frac{\mathscr{C}(n_j, \ell_j; \sigma)}{\sigma^{\ell_j}},$$

where $\mathscr{V}_{n,k}$ is in (6) and $\mathscr{C}(n, k; \sigma) := (k!)^{-1} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n$ for $k \leq n$ are the generalized factorial coefficients, which can be computed by exploiting recursive formulas. Hence, the EPPF of a NGG multinomial process has a simpler form compared to the general equation (11), because it only depends on the integers $\ell_1, \ldots, \ell_k$. This also enables the actual evaluation of the upper bound in Theorem 3, thus yielding

$$\frac{H!}{H^k(H-k)!} \leq \frac{\Pi_H(n_1, \ldots, n_k)}{\Pi_\infty(n_1, \ldots, n_k)} \leq \sum_{\boldsymbol{\ell}} \frac{\mathscr{V}_{n,|\boldsymbol{\ell}|}}{\mathscr{V}_{n,k}} \left(\frac{c}{\sigma H}\right)^{|\boldsymbol{\ell}|-k} \prod_{j=1}^{k} \frac{\mathscr{C}(n_j, \ell_j; \sigma)}{\mathscr{C}(n_j, 1; \sigma)},$$

where $\mathscr{C}(n_j, 1; \sigma) = \sigma(1-\sigma)_{n_j-1}$.

The availability of Π_H naturally leads one to address the problem of determining the distribution of the number of partition sets $K_{n,H}$, that is, the law of the number of distinct values observed in a sample $\boldsymbol{\theta}^{(n)}$ from a NIDM process prior. As one might expect, $K_{n,H}$ converges to the number of partition sets $K_{n,\infty}$, namely the number of distinct values generated by an exchangeable n-sample from homogeneous NRMI, when $H \to \infty$. Another interesting connection between $K_{n,H}$ and $K_{n,\infty}$ is formalized in the following theorem.

Theorem 4. *For any* $k \leq \min\{H, n\}$

$$\mathbb{P}(K_{n,H} = k) = \frac{H!}{H^k(H-k)!} \sum_{\ell=0}^{n-k} \frac{1}{H^\ell} \mathscr{S}(\ell + k, k) \mathbb{P}(K_{n,\infty} = \ell + k),$$

where $\mathscr{S}(\ell, k) = (k!)^{-1} \sum_{r=0}^{k} (-1)^{k-r} \binom{k}{r} r^\ell$ *is the Stirling number of the second kind for* $\ell, k \geq$

0. *Moreover, the expected value of* $K_{n,H}$ *is given by*

$$\mathbb{E}(K_{n,H}) = H - H\mathbb{E}\left(\left(\sum_{h=1}^{H-1}\pi_h\right)^n\right) = H - H\mathbb{E}\left(\left(1-\frac{1}{H}\right)^{K_{n,\infty}}\right).$$

From Theorem 4 it can be easily seen that $\mathbb{P}(K_{n,H}=k) \to \mathbb{P}(K_{n,\infty}=k)$ for all k as $H\to\infty$, since $\mathscr{S}(k,k)=1$. Hence, $K_{n,H} \xrightarrow{d} K_{n,\infty}$. Additionally, we have that $\mathbb{E}(K_{n,H}) \to \mathbb{E}(K_{n,\infty})$ as $H\to\infty$, and the following asymptotic expansion holds $\mathbb{E}(K_{n,H})/\mathbb{E}(K_{n,\infty}) = 1-(2H)^{-1}\left(\mathbb{E}(K_{n,\infty}^2)/\mathbb{E}(K_{n,\infty})-1\right)+O\left(H^{-2}\right)$, as $H\to\infty$. Thus, the convergence of the expected value to the infinite case occurs at the linear rate $O(1/H)$. We expanded the above ratio up the the second order, to gain further understanding about the speed at which $\mathbb{E}(K_{n,H})$ approaches its limit: broadly speaking, quick convergence occurs whenever $\mathbb{E}(K_{n,\infty}^2) \approx \mathbb{E}(K_{n,\infty})$, which is the case when $\mathbb{E}(K_{n,\infty})$ is relatively small. On the other hand, one needs a large value of H when trying to approximate an infinite-dimensional process having a high number of expected clusters in a sample of size n. This is in line with the discussion of Example 1.

Example 3 (Dirichlet multinomial process, cont'd)**.** A straightforward application of Theorem 4 yields $\mathbb{P}(K_{n,H}=k) = [H!/(H-k)!](-1)^k\mathscr{C}(n,k;-c/H)/(c)_n$, for $k=1,\dots,\min\{H,n\}$. This simple form is due to the Gibbs-type structure of the Dirichlet multinomial process, and indeed the above formula might be deduced from Gnedin and Pitman (2005). In the relevant particular case $c=H$, that is, when the weights of the NIDM process are uniformly distributed, the following simplification occurs: $\mathbb{P}(K_{n,H}=k) = n!\binom{H}{k}\binom{n-1}{k-1}/(H)_n$, for $k=1,\dots,\min\{H,n\}$, a particular instance of hypergeometric distribution. As for the expected value of $K_{n,H}$, for any value of c we have the following simple formula: $\mathbb{E}(K_{n,H}) = H-(H-1)(c+1-c/H)_{n-1}/(c+1)_{n-1}$. Note that, as $H\to\infty$, the right-hand-side converges to the expected number of cluster induced by the Dirichlet process with concentration parameter c.

Example 4 (NGG multinomial process, cont'd)**.** Direct application of Theorem 4 yields

$$\mathbb{P}(K_{n,H}=k) = \frac{H!}{(\sigma H)^k(H-k)!}\sum_{\ell=0}^{n-k}\frac{\mathscr{V}_{n,\ell+k}}{H^\ell}\mathscr{S}(\ell+k,k)\frac{\mathscr{C}(n,\ell+k;\sigma)}{\sigma^\ell},$$

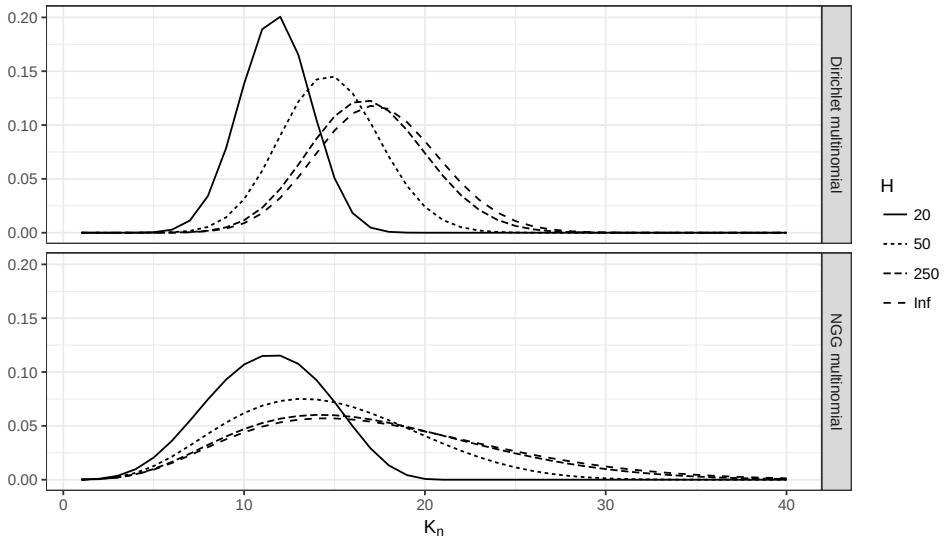

Figure 1: Distribution of $K_{100,H}$ with $n = 100$, under a Dirichlet multinomial process ($c = 5.87$), and a NGG multinomial process ($c = 1/2, \kappa = 1, \sigma = 1/2$), for different levels of $H \in \{20, 50, 250, \infty\}$. The distribution is depicted only for the values in the interval $[1, 40]$ for graphical reasons.

for any $k = 1, \ldots, \min\{H, n\}$, since $\mathbb{P}(K_{n,\infty} = k) = \mathscr{V}_{n,k}\sigma^{-k}\mathscr{C}(n, k; \sigma)$. Furthermore, the expected value of $K_{n,H}$ is given by $\mathbb{E}(K_{n,H}) = H - H\sum_{\ell=1}^{n}(1 - 1/H)^{\ell}\mathscr{V}_{n,\ell}\mathscr{C}(n, \ell; \sigma)\sigma^{-\ell}$.

To illustrate the clustering mechanism, in Figure 1 we depicted the distribution of the random variable $K_{100,H}$ under both a Dirichlet multinomial process and a NGG multinomial process, for different values of H. To make these prior choices comparable, we have set the hyperparameters c and (β, σ) such that the expected number of clusters for the corresponding infinite-dimensional NRMIs $\mathbb{E}(K_{100,\infty})$ is the same.

As highlighted in Figure 1, the distribution of $K_{100,H}$ under the NGG multinomial prior is "flatter", i.e. less informative, compared to the one induced by the Dirichlet multinomial prior, for any of the values of H being considered. We note that the variance of $K_{n,H}$ in the NGG prior can be tuned through the parameter σ. When $\sigma \to 0$ the Dirichlet multinomial case is recovered. If $H \to \infty$, this effect of σ was extensively discussed in Lijoi et al. (2007) and it comes as no surprise it is reflected also in the finite H case, in view of Theorem 4. Theorems 2-4 show that, despite the technical difficulties posed by the finite H setting, distributional properties similar to those available in the infinite-dimensional case can still

be achieved. Nonetheless, we remark that if one is interested in approximating the infinite-dimensional NRMI prior, a relatively large value of H is typically required. For instance, Figure 1 suggests that we should set $H = 250$ to satisfactorily approximate both the Dirichlet and the NGG processes with their NIDM counterparts, in this specific setting.

4 Posterior characterizations

We complete here the picture of the distributional properties of NIDM processes by determining their posterior distribution. We will refer to a framework where

$$(\theta_1, \ldots, \theta_n \mid \tilde{p}_H) \overset{\text{iid}}{\sim} \tilde{p}_H, \qquad \tilde{p}_H \sim \text{NIDM}(c, \rho; P). \tag{12}$$

and will provide two representations of the posterior distribution of $\tilde{p}_H$ one of which is effectively described in terms of the the multiroom Chinese restaurant metaphor introduced in Camerlenghi et al. (2018). While our results are general, particular emphasis is given to the NGG multinomial process and, despite the lack of conjugacy, we are able to devise algorithms for exact (iid) sampling of the posterior. In contrast with most widely known samplers for homogeneous NRMIs (e.g. Lijoi and Prünster, 2010; Barrios et al., 2013; Favaro and Teh, 2013; Arbel and Prünster, 2017), which either require Gibbs-sampling strategies or truncating certain series representations, the posterior laws of NIDM processes can be sampled exactly.

4.1 Predictive distributions and posterior laws

Since the EPPF Π_H can be evaluated through Theorem 2, it is straightforward to compute the predictive distributions. To this end, for any $n \geq 1$ we define a positive random variable U whose density function on $\mathbb{R}^+$, conditional on the sample $\boldsymbol{\theta}^{(n)}$, equals

$$q_H(u) \propto u^{n-1} \mathrm{e}^{-c\psi(u)} \prod_{j=1}^{k} \Delta_{n_j,H}(u). \tag{13}$$

The normalizing constant of the above density is finite and it essentially identifies the EPPF of a NIDM process. The density function q_H parallels the one, say q_∞, of the latent variable

appearing in the posterior representation for NRMIs. See James et al. (2009). Note that q_H converges pointwise to q_∞ as $H \to \infty$ since $\Delta_{m,H}(u) \to \tau_m(u)$ for any $m \geq 1$ and $u > 0$.

Corollary 2. *Let $\theta_1, \ldots, \theta_n$ be as in (12), and such that they admit k distinct values $\theta_1^*, \ldots, \theta_k^*$ with θ_j^* having frequency n_j. Then $\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}^{(n)}) = w_0(\boldsymbol{n}^{(k)})P(A) + \sum_{j=1}^k w_j(\boldsymbol{n}^{(k)})\delta_{\theta_j^*}(A)$, where $\boldsymbol{n}^{(k)} = (n_1, \ldots, n_k)$ and for any $j = 1, \ldots, k$*

$$w_0(\boldsymbol{n}^{(k)}) = \left(1 - \frac{k}{H}\right)\frac{c}{n}\int_{\mathbb{R}^+} u\Delta_{1,H}(u)q_H(u)\mathrm{d}u, \quad w_j(\boldsymbol{n}^{(k)}) = \frac{1}{n}\int_{\mathbb{R}^+} u\frac{\Delta_{n_j+1,H}(u)}{\Delta_{n_j,H}(u)}q_H(u)\mathrm{d}u.$$

This result is reminiscent of the predictive distributions obtained for homogeneous NRMIs, and indeed similar sampling strategies can be borrowed from that context. For example, note that conditionally on the latent variable U one has $\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}^{(n)}, U) \propto (1 - k/H)\, c\Delta_{1,H}(U)P(A) + \sum_{j=1}^k \{\Delta_{n_j+1,H}(U)/\Delta_{n_j,H}(U)\}\delta_{\theta_j^*}(A)$. Hence, one can devise a generalized Pólya-urn scheme by first drawing U from its density q_H and then sampling from the predictive distribution using the above formula. The terms $\Delta_{m,H}$ might be expensive to compute in practice, mainly because of their combinatorial nature. However, this issue can be attenuated by relying on the recursive definition of $\Delta_{m,H}$ provided in James et al. (2006). Furthermore, in the fairly general NGG multinomial case, the weights $\Delta_{m,H}$ have an explicit formula, and this allows for the implementation of an exact sampling algorithm for U; see the supplementary material.

Remark 1. If one is only interested in obtaining an exchangeable draw for $\boldsymbol{\theta}^{(n)}$, a direct strategy consists in simulating $\tilde{p}_H$ and then, conditionally on it, sampling iid observations from $\tilde{p}_H$, according to (12). Indeed, any ID random variable whose Laplace exponent is available in closed form can be sampled, for example through the general algorithm of Ridout (2009), and this enables the simulation of NIDM processes.

We now provide a posterior characterization for the law of the random measure $\tilde{\mu}_H$ given $\boldsymbol{\theta}^{(n)}$: the posterior distribution of $\tilde{p}_H$ can, then, be recovered by normalization. To ease notation, the posterior law is expressed conditionally on the latent variable U, whose density is q_H. Moreover, when the sample $\boldsymbol{\theta}^{(n)} = (\theta_1, \ldots, \theta_n)$ displays $k < H$ distinct values $\theta_1^*, \ldots, \theta_k^*$, we let $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ represent the point masses in $\tilde{p}_H$ that are not included in $\boldsymbol{\theta}^{(n)}$, corresponding to the original $\tilde{\theta}_h$'s, not included in $\boldsymbol{\theta}^{(n)}$, up to a permutation.

Theorem 5. *Let $\theta_1, \ldots, \theta_n$ be as in (12) and, conditionally on $\boldsymbol{\theta}^{(n)}$, let U be a random variable with density q_H as in (13). Then*

$$(\tilde{\mu}_H \mid \boldsymbol{\theta}^{(n)}, U) \stackrel{d}{=} \sum_{j=k+1}^{H} J_j^* \delta_{\bar{\theta}_j} + \sum_{j=1}^{k} (J_j^* + I_j) \delta_{\theta_j^*}, \tag{14}$$

where $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ are iid draws from P and $(J_h^* \mid U) \stackrel{\text{iid}}{\sim} \text{ID}\left(\frac{c}{H}, \rho^*\right)$, with $\rho^*(s) = \mathrm{e}^{-Us}\rho(s)$, for $h = 1, \ldots, H$. Finally, the jumps $I_1, \ldots, I_k$ are independent and nonnegative random variables characterized by $\mathbb{E}\left(\mathrm{e}^{-\lambda I_j} \,\middle|\, \boldsymbol{\theta}^{(n)}, U\right) = \Delta_{n_j,H}(\lambda + U)/\Delta_{n_j,H}(U)$, for $j = 1, \ldots, k$.

This representation is related to the posterior representation for homogeneous NRMIs, which can be recovered as $H \to \infty$. Indeed, the first term in (14) converges to a CRM with the exponentially tilted Lévy intensity ρ^*, as a consequence of Theorem 1. On the other hand, $J_j^* \xrightarrow{d} 0$ and $\mathbb{E}\left(\mathrm{e}^{-\lambda I_j} \mid \boldsymbol{\theta}^{(n)}, U\right) \to \tau_{n_j}(\lambda + U)/\tau_{n_j}(U)$ for any $j = 1, \ldots, k$: hence, the second term on the right-hand-side of (14) converges to the fixed jumps component of the posterior representation of NRMIs. See James et al. (2009). Interestingly, a structural property is shared by NRMIs and NIDMs: conditionally on a latent variable U the posterior law is a mixture of: (i) a component with a tilted intensity and (ii) a collection of independent jumps corresponding to the distinct values $\theta_1^*, \ldots, \theta_k^*$ in the sample $\boldsymbol{\theta}^{(n)}$. However, it must be stressed that for NIDMs the tilted component vanishes as soon as $k = H$ distinct values are recorded in the sample, and the posterior distribution will coincide with the law of a measure with jumps at fixed locations identified by the distinct values $\theta_1^*, \ldots, \theta_H^*$.

Example 5 (Dirichlet multinomial process, cont'd)**.** Note that $\Delta_{m,H}(\lambda + u)/\Delta_{m,H}(u) = \tau_m(\lambda + u)/\tau_m(u)$ for any $m \geq 1$ and H, implying that the random variables in Theorem 5 have a simple form $(J_j^* \mid \boldsymbol{\theta}^{(n)}, U) \stackrel{\text{iid}}{\sim} \text{GAMMA}\,(c/H, 1 + U)$, $(I_j \mid \boldsymbol{\theta}^{(n)}, U) \stackrel{\text{ind}}{\sim} \text{GAMMA}(n_j, 1 + U)$, for $j = 1, \ldots, H$, where we agree that $I_j = 0$ a.s. for any $j > k$. Hence, after normalization, we get the usual Dirichlet structure with updated parameters, which does not depend on the latent variable U. Finally, it is easy to check that Corollary 2 can be specialized to obtain the well-known predictive distributions of the Dirichlet multinomial process.

Example 6 (NGG multinomial process, cont'd)**.** For any $m \geq 1$ one has

$$\Delta_{m,H}(u) = \sum_{\ell=1}^{m} \left(\frac{c}{H}\right)^{\ell-1} \frac{\mathscr{C}(m, \ell; \sigma)}{\sigma^\ell} (\kappa + u)^{-m+\ell\sigma}.$$

Moreover, the random variables $J_1^*, \ldots, J_H^*$ of Theorem 5 are *conditionally conjugate*, because $\rho^*(s) = \mathrm{e}^{-Us}\rho(s) = \Gamma(1-\sigma)^{-1}s^{-1-\sigma}\mathrm{e}^{-(\kappa+U)s}$ identifies an updated generalized gamma process. The distribution of each $J_1^*, \ldots, J_H^*$ is known as *tempered stable* and there exists several methods for drawing samples from it; see e.g. Ridout (2009) and references therein. Furthermore, the random variables $I_1, \ldots, I_k$, given U, have the following conditional mixture densities

$$f_{I_j}(w \mid u) = \sum_{\ell=1}^{n_j} \frac{\xi_{n_j,\ell,H}(u)}{\Delta_{n_j,H}(u)} \textsc{gamma}(w; n_j - \ell\sigma, \kappa + u), \tag{15}$$

for $j = 1, \ldots, k$, where $\textsc{gamma}(w; a, b)$ denotes the density function of a gamma random variable. Finally, for any $A \in \mathscr{B}(\Theta)$ some algebra yields

$$\begin{aligned}\mathbb{P}(\theta_{n+1} \in A \mid \boldsymbol{\theta}^{(n)}, U) \propto &\left(1 - \frac{k}{H}\right) c(\kappa + U)^{\sigma} P(A) \\ &+ \sum_{j=1}^{k} \left[\frac{1}{H} c(\kappa + U)^{\sigma} + \sum_{\ell_j=1}^{n_j} \frac{\xi_{n_j,\ell_j,H}(U)}{\Delta_{n_j,H}(U)} (n_j - \ell_j\sigma)\right] \delta_{\theta_j^*}(A).\end{aligned} \tag{16}$$

Remark 2. To enable posterior inference through random sampling it suffices to simulate iid U values from q_H and, then, make use of the above posterior representation. Although q_H is known up to a normalizing constant, we can nonetheless draw samples by acceptance-rejection algorithms. The simulation of the limiting density q_∞ was typically addressed via Markov Chain Monte Carlo (Lijoi et al., 2007). However, this further complication can be avoided in the NGG setting, given the availability of algorithms for exact sampling, which are discussed in the supplementary material both for q_H and q_∞. As such, these algorithms might be useful beyond their application to NIDMs. This data-augmentation scheme circumvents the need of computing the weights $\mathscr{V}_{n,k}$, which may be numerically unstable for large values of n (Lijoi et al., 2007). In contrast, a potential bottleneck of such a strategy is the evaluation of the generalized factorial coefficients $\mathscr{C}(n, k; \sigma)$ when n is large, although the availability of recursive formulas for $\mathscr{C}(n, k; \sigma)$ may mitigate the issue.

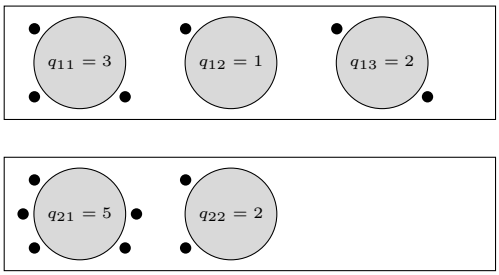

Figure 2: The multiroom chinese restaurant metaphor: circles represent tables and bullets represent customers. The number of tables for these two rooms are $(\ell_1, \ell_2) = (3, 2)$ so that $|\boldsymbol{\ell}| = \ell_1 + \ell_2 = 5$. The number of customers eating the first dish (i.e., first room) is $n_1 = \sum_{r=1}^{\ell_1} q_{1r} = 6$, while the number of customers eating the second dish (i.e., second room) are $n_2 = \sum_{r=1}^{\ell_2} q_{2r} = 7$.

4.2 Multiroom Chinese restaurant metaphor

To gain further insights about structural properties of NIDM processes, we now describe a data-augmentation based on the hierarchical representations (10)-(11). To this purpose, it is worth recalling the so-called *multiroom Chinese restaurant* metaphor coined by Camerlenghi et al. (2018), which can be adapted to NIDM processes. Suppose that there exists a restaurant which serves a *finite* number of dishes H, corresponding to iid draws from the diffuse P. The restaurant has infinitely many rooms, and each room contains infinitely many tables and is associated to a single dish out of the H available from the menu. The first customer seats at one of the tables of the first room and selects a dish. The nth customer can either select a dish previously chosen by the other $n-1$ customers or she can choose a new dish. In the former case, she will be seated in the room serving the dish of choice and she may be seated either at a new table or at an existing one. If a new dish is chosen, she will sit in a new room and at a new table. An illustration of this generative scheme is depicted in Figure 2.

Recalling the notation used so far, the entries of $\boldsymbol{\theta}^{(n)} = (\theta_1, \ldots, \theta_n)$ represent the dishes eaten by the n customers of the restaurant, whereas the labels identifying the single tables and their respective frequencies tables are unobservable and, hence, *latent quantities*. Specifically, we consider the latent random variables $\boldsymbol{T}^{(n)} = (T_1, \ldots, T_n)$, which can be thought of as the label of the table where each customer is seated. Recall that k distinct dishes are served at the restaurant, that is, there are $\theta_1^*, \ldots, \theta_k^*$ distinct values having frequencies $n_1, \ldots, n_k$, meaning that the total number of customers seating in room j or, equivalently, eating dish

j, corresponds to the frequency n_j. Finally, each $\ell_j \in \{1, \dots, n_j\}$ represents the number of tables in room j where customers are seated, while each q_{jr} denotes the number of customers seating at table r in room j, so that $\sum_{j=1}^{k} \sum_{r=1}^{\ell_j} q_{jr} = \sum_{j=1}^{k} n_j = n$. When $H \to \infty$, the probability of observing a new table where the same dish is being served tends to zero, since $\ell_j \longrightarrow 1$ for any $j = 1, \dots, k$. This in turn implies $|\boldsymbol{\ell}| \longrightarrow k$ and, hence, each room will have only one table. We denote the $|\boldsymbol{\ell}|$ distinct labels of the tables with $T^*_{j1}, \dots, T^*_{j\ell_j}$ having frequencies q_{jr} for $r = 1, \dots, \ell_j$ and $j = 1, \dots, k$. Thus, the joint augmented model for $\boldsymbol{\theta}^{(n)} = (\theta_1, \dots, \theta_n)$ and $\boldsymbol{T}^{(n)} = (T_1, \dots, T_n)$ follows immediately from equation (11). Indeed, one has the following

Corollary 3. *The joint probability distribution of* $(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}, \Psi_{n,H})$*, where* $\Psi_{n,H}$ *is the partition of* $[n]$ *induced by* $\boldsymbol{\theta}^{(n)}$ *through* (12)*, is*

$$\prod_{j=1}^{k} P(\mathrm{d}\theta^*_j) \prod_{r=1}^{\ell_j} P(\mathrm{d}T^*_{jr}) \times \left[\frac{H!}{(H-k)!} \frac{1}{H^{|\boldsymbol{\ell}|}} \Pi_\infty(q_{11}, \dots, q_{1\ell_1}, \dots, q_{k1}, \dots, q_{k\ell_k}) \right]. \tag{17}$$

The baseline distribution of $\boldsymbol{T}^{(n)}$ is set equal to P for simplicity, but any other diffuse probability measure would obviously work. Indeed, only the clustering mechanism implied by the table configuration is relevant for our purposes, and not the actual labels. The representation in (17) thus suggests that, conditionally on the table configuration induced by $\boldsymbol{T}^{(n)}$, the predictive distribution for θ_{n+1} can be easily obtained. Moreover, given the previously observed values $\boldsymbol{\theta}^{(n)}$, the table configuration can be drawn through a Gibbs sampler. For the sake of the exposition, we do not attempt the full derivation of these conditional distributions, but the interested reader may refer to Camerlenghi et al. (2018) for a detailed discussion, though in a different setting.

Example 7 (NGG multinomial process, cont'd)**.** Conditionally on the latent random variables $\boldsymbol{T}^{(n)}$, the predictive probabilities can be readily obtained from equation (17), so that

$$\begin{aligned} \mathbb{P}(\theta_{n+1} \in A \,|\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}) = \\ = \left(1 - \frac{k}{H}\right) \frac{\mathscr{V}_{n+1,|\boldsymbol{\ell}|+1}}{\mathscr{V}_{n,|\boldsymbol{\ell}|}} P(A) + \sum_{j=1}^{k} \left[\frac{1}{H} \frac{\mathscr{V}_{n+1,|\boldsymbol{\ell}|+1}}{\mathscr{V}_{n,|\boldsymbol{\ell}|}} + \frac{\mathscr{V}_{n+1,|\boldsymbol{\ell}|}}{\mathscr{V}_{n,|\boldsymbol{\ell}|}} (n_j - \ell_j \sigma) \right] \delta_{\theta^*_j}(A). \end{aligned}$$

Hence, a relevant simplification occurs when considering NGG multinomial processes. In par-

ticular, the above conditional distribution depends on the table configuration only through the number of distinct tables $\ell_1, \ldots, \ell_k$ rather than on the table-specific frequencies q_{jr}. This is a major computational advantage, since we only need to sample k latent variables rather than n, as in general NIDM processes. Moreover, the above conditional law is intimately related to (16), as it arises from the combination of an augmentation over the table configuration and of a marginalization with respect to the latent variable U.

In the following, we expand the posterior characterization of Theorem 5 by conditioning also on the table configuration. The random variable U, conditionally on $(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)})$, is a nonnegative latent variable whose density function is given by

$$q_\infty(u) \propto u^{n-1} \mathrm{e}^{-c\psi(u)} \prod_{j=1}^{k} \prod_{r=1}^{\ell_j} \tau_{q_{jr}}(u). \tag{18}$$

Thus, conditionally also on $\boldsymbol{T}^{(n)}$, the latent variable U has the same structure of that involved in the posterior derivation of NRMIs. Similar simplifications occur also for the fixed jump component, as summarized in the next corollary.

Corollary 4. *Let $\theta_1, \ldots, \theta_n$ be a draw from an exchangeable sequence directed by $\tilde{p}_H \sim \mathrm{NIDM}(c, \rho; P)$ as in (12). Moreover, let the conditional distribution of U, given $(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)})$, have density function q_∞ defined as in (18). Then,*

$$(\tilde{\mu}_H \mid \boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}, U) \stackrel{d}{=} \sum_{j=k+1}^{H} J_j^* \delta_{\bar{\theta}_j} + \sum_{j=1}^{k} (J_j^* + \sum_{r=1}^{\ell_j} I_{jr}) \delta_{\theta_j^*}(A),$$

where $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ are iid draws from P, and $(J_h^* \mid U) \stackrel{\text{iid}}{\sim} \mathrm{ID}\,(c/H, \rho^*)$, with $\rho^*(s) = \mathrm{e}^{-Us}\rho(s)$, for $h = 1, \ldots, H$. Moreover, the jumps I_{jr} are independent and nonnegative random variables having density $f_{jr}(s \mid \boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}, U) \propto \mathrm{e}^{-sU} s^{q_{jr}} \rho(s)$, for $r = 1, \ldots, \ell_j, j = 1, \ldots, k$.

Hence, conditionally on the table configuration, the posterior structure of $\tilde{p}_H$ closely resemble the one of homogeneous NRMIs, for any finite value of H. Specifically, the distribution of the random variables I_{jr} have the same distributional structure of the jumps in NRMI–based models.

Example 8 (NGG multinomial process, cont'd)**.** Specializing Corollary 4 we obtain that $(\sum_{r=1}^{\ell_j} I_{jr} \mid \boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}, U) \sim \text{GAMMA}(n_j - \ell_j\sigma, \kappa + U)$, for $j = 1, \ldots, k$, which depends on $\boldsymbol{T}^{(n)}$ only through the number of tables $\ell_1, \ldots, \ell_k$. Note that this representation is essentially the augmentation of equation (15) with respect to the number of tables. Note that when $\sigma = 0$, the posterior law $\tilde{p}_H$ becomes independent on the table configuration.

5 The INVALSI dataset

We consider a publicly available dataset gathered by INVALSI institute, which is a public research center for the assessment of the Italian education system. In particular, the 2016-2017 dataset we are going to analyze is part of a national examination program conducted in Italy with the aim of "carrying out periodic and systematic checks on knowledge and skills of students", as declared in the official documentation of the INVALSI statistical service[1]. A great effort was put by the INVALSI in order to quantify the *added-value* of a school, based on these data. The Bayesian framework is a natural choice when trying to combine multiple sources of information, i.e. the schools. This can be accomplished through hierarchical nonparametric models, which enable flexible borrowing of information between different institutions. A broad and systematic socio-demographic analysis is beyond the scope of this paper and, hence, we focus on the presentation of novel modeling strategies based on NIDM priors.

We focus on data related to 8th grade students from schools in the city of Bologna: more specifically we consider those questions related to the comprehension of the Italian language. Having omitted few observations for which covariates were not available, the resulting dataset comprises a total of $N = 8126$ observations (students), belonging to 84 educational institutions. The INVALSI test has 45 questions and the performance of each student might be well summarized by the proportions of correct answers. To ease the modeling process we take a logistic transformation of the original proportions, and we define the score S_{ij} for the ith student in the jth school as

$$S_{ij} := \text{logit}\left(\frac{\#\text{ of correct answers, } i\text{th student } j\text{th school} + 1/2}{\#\text{ of questions} + 1}\right),$$

for $i = 1, \ldots, N_j$ and $j = 1, \ldots, 84$, where N_j denotes the number of students in the jth

[1]The documentation (in Italian) is available at: https://invalsi-serviziostatistico.cineca.it

school. In the above ratio, we added a small correction to the original proportions to avoid boundary issues. Such a transformation maps the original scores into $\mathbb{R}$, and therefore it is more amenable for classical linear modeling with Gaussian errors. Consistently with this, we model the scores as follows

$$S_{ij} = \mu_j + \boldsymbol{z}_{ij}^{\intercal}\boldsymbol{\gamma} + \epsilon_{ij}, \qquad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2),$$

for $i = 1, \ldots, N_j$ and $j = 1, \ldots, 84$, where $\boldsymbol{z}_{ij} = (1, z_{ij1}, \ldots, z_{ijp})^{\intercal}$ is a vector of student-specific covariates which are associated to a $(p+1)$–dimensional vector of regression coefficients $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_p)^{\intercal}$. Each vector $\boldsymbol{z}_{ij}$ encodes student-specific categorical variables, namely: the gender of the student, the education level of her/his father and mother (primary school, secondary school, etc.), the employment status of her/his father and mother, the regularity of the student (i.e. regular, in late, etc.), and the citizenship (Italian, first generation immigrant, etc.). Moreover, the coefficients $\mu_1, \ldots, \mu_{84}$ represents the school effects or, in the terminology used so far, the *added-value* of the school given a set of covariates, thus being the main quantity of interest in our analysis. Note that since the intercept term is included in $\boldsymbol{z}_{ij}$, the coefficients $\mu_1, \ldots, \mu_{84}$ are not identified. In practice, this is not a concern if inference is based on the "centred" set of parameters $\eta_j = \gamma_0 + \mu_j$, for $j = 1, \ldots, 84$, rather than the original random effects.

We aim at introducing a flexible nonparametric prior, for the school effects $\mu_1, \ldots, \mu_{84}$, that allows for: (i) borrowing information across schools; (ii) arbitrary deviations from Gaussianity, and (iii) robustness under model misspecifications. Gaussian mixture models are able to capture all these three aspects and we specifically let

$$(\mu_j \mid \mathcal{P}) \overset{\text{iid}}{\sim} \mathcal{P}, \qquad \mathcal{P} = \sum_{h=1}^{H} \pi_h \mathcal{N}(\bar{\mu}_h, \bar{\sigma}_h^2), \tag{19}$$

for $j = 1, \ldots, 84$. Selecting the appropriate number of mixture components H is typically a difficult task, and the BIC index — customarily employed in this framework — is not theoretically well justified here. Hence, overfitted mixture models can be exploited to circumvent this issue. Here we propose the usage of general NIDM processes, which amounts to have the

following prior specification for the parameters in (19)

$$(\pi_1, \ldots, \pi_{H-1}) \sim \text{NID}\left(\frac{c}{H}, \ldots, \frac{c}{H}; \rho\right), \qquad (\bar{\mu}_h, \bar{\sigma}_h^2) \overset{\text{iid}}{\sim} P, \qquad h = 1, \ldots, H,$$

where P is a diffuse probability measure on $\mathbb{R} \times \mathbb{R}^+$. By enlarging the class of priors from the Dirichlet multinomial to general NIDM multinomial processes we are essentially acting on the robustness requirement, that is, we are ensuring that the clustering mechanism is less affected by specific choices of the total mass c, whose specification is often critical.

For this specific application, we employed in (19) a NGG multinomial process having jump intensity (3). The hyperparameters are set equal to $c = 0.1, \kappa = 0.1$ and $\sigma = 0.87$ and combined with a very conservative upper bound $H = 70$ for the number of mixture components. The a priori effect of these values on the number of cluster $K_{n,H}$ is depicted in Figure 3, where it is compared with the distribution induced by a Dirichlet multinomial process ($c = 45.21$), a Dirichlet process ($c = 21.69$), and a NGG process ($c = 0.2, \kappa = 0.1$, $\sigma = 0.8$), having roughly the same expected value $\mathbb{E}(K_{n,H}) \approx 35$. As apparent from Figure 3, the parameter σ plays a crucial role in controlling the variability of $K_{n,H}$. Hence, one can obtain flat distributions for $K_{n,H}$, which leads to more robust inference on the cluster configurations. Such an effect persists a posteriori, as we shall discuss later on and is further corroborated by the extensive simulation study provided in the supplementary material. As for the choice of P, we assume the conditionally conjugate prior $\bar{\mu}_h \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\bar{\mu}}^2)$, $\bar{\sigma}_h^{-2} \overset{\text{iid}}{\sim} \text{GAMMA}(a_{\bar{\sigma}}, b_{\bar{\sigma}})$, for $h = 1, \ldots, H$, where we set $\sigma_{\bar{\mu}}^2 = 0.1$ and $a_{\bar{\sigma}} = 1.5$, $b_{\bar{\sigma}} = 0.045$. It is well-known that the choice of these parameters is delicate, since it may influence posterior inference. This issue however affects both the Dirichlet and NGG specifications. To conclude our Bayesian formulation we consider a multivariate Gaussian prior for the regression coefficients $\boldsymbol{\gamma}$, and an inverse Gamma prior for the residuals variance σ_ϵ^2, and we let $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{B})$, $\sigma_\epsilon^{-2} \sim \text{GAMMA}(a_\sigma, b_\sigma)$, where we set $\boldsymbol{b} = 0$ and $\boldsymbol{B} = \text{diag}(100, \ldots, 100)$, to incorporate the neutral hypothesis of no relevant effects and $a_\sigma = b_\sigma = 1$, which induces a fairly non-informative prior. Posterior inference was conducted through a Gibbs sampling, whose details can be found in the supplementary material. For comparison, we also estimated the same model under a Dirichlet multinomial, a Dirichlet process, and a NGG prior as for Figure 3. We run the algorithm for 20′000 iterations after a burn-in period of 2′000 simulations; the traceplots show no evidence against convergence and an excellent mixing.

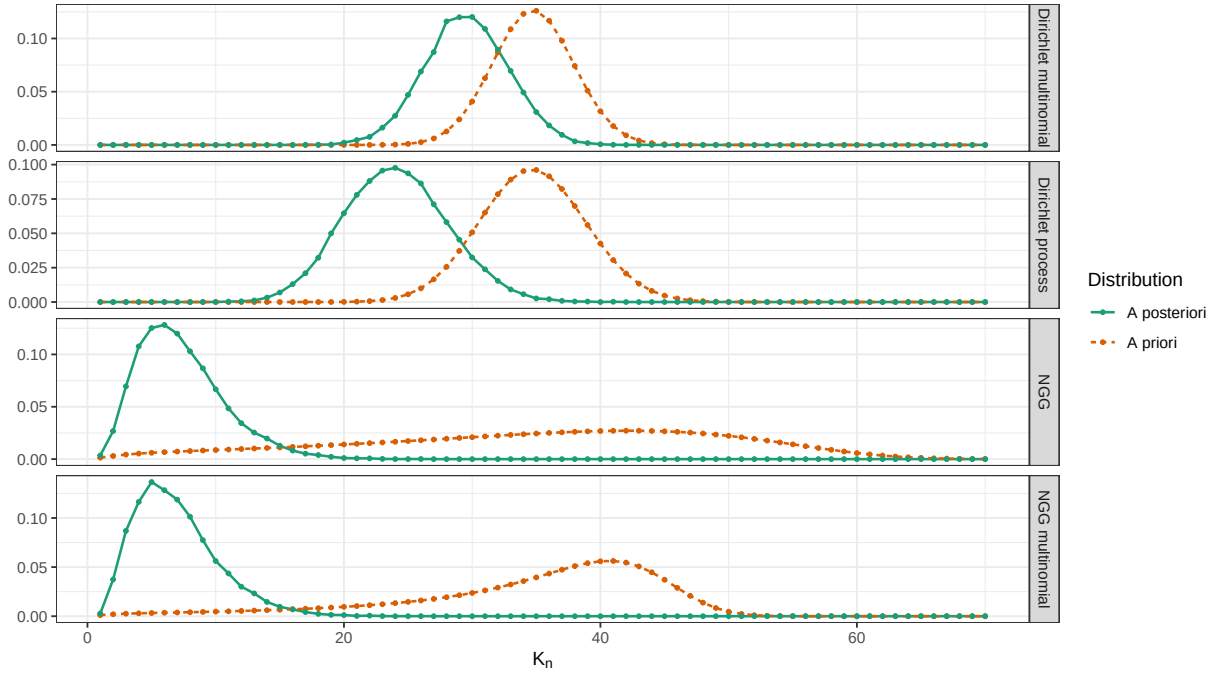

Figure 3: A priori and a posteriori distribution of the number of cluster $K_{n,H}$ in the INVALSI application, under a Dirichlet multinomial prior, a Dirichlet process, a NGG multinomial process prior, and NGG process with $H = 70$ and $\mathbb{E}(K_{n,H}) \approx 35$.

From Figure 3, it is apparent that the posterior of $K_{n,H}$ is heavily influenced by the choice of the prior. In the Dirichlet multinomial and Dirichlet process cases, the number of mixture components peaks at around 29 and 24, respectively, as a consequence of the strongly informative prior distribution. Conversely, when the less informative NGG priors are used, the data adaptively select a much smaller number of components, peaking at around 4-6 mixture components. As expected, the infinite-dimensional NGG and the NGG multinomial behave in a similar fashion. Thus, if a simple Gaussian random effect model were employed, the random effects η_j would be overshrunk towards the global mean, potentially affecting the quality of the analysis. On the other hand, the large number of clusters induced by Dirichlet priors could lead to undershrinkage, resulting in a model that is overly adapted to the data.

To empirically validate our expectations, we estimated a Gaussian random effects model, corresponding to $H = 1$, and a so-called "no pooling" model, in which no shrinkage is induced on the school effects η_j. All these models were compared in terms of the DIC, WAIC

	DIC	WAIC	LOO	p_{DIC}	p_{WAIC}	p_{LOO}
Dirichlet multinomial	18960.33	18957.71	18961.54	109.65	107.04	108.95
Dirichlet process	18957.96	18955.49	18959.21	106.97	104.51	106.36
NGG multinomial	18955.98	18955.98	18957.62	96.77	95.14	96.78
NGG	18955.96	18954.25	18957.59	97.28	95.57	97.24
Gaussian	18967.82	18966.49	18969.31	95.62	94.28	95.70
No pooling	18975.47	18971.92	18976.50	120.90	117.35	119.64

Table 1: Information criterion DIC, WAIC and LOO (deviance scale, lower is better) in the INVALSI application, defined as in Gelman et al. (2014). The quantities p_{DIC}, p_{WAIC}, and p_{LOO} are the associated estimates for the "effective number of parameters".

(Gelman et al., 2014), and the Bayesian leave-one-out (LOO), computed as in Vehtari et al. (2017). These indexes, reported in Table 1, represent an overall goodness of fit measurement and yield, as a byproduct, an estimate for the "effective number of parameters", which is a measure of the model complexity. In addition, in Table 2 we report the posterior mean of some (centred) school effects $\eta_j - \gamma_0$. According to the DIC, the WAIC, and the LOO it is clear that a certain amount of shrinkage is required: the "no pooling" model is overly complex. Conversely, the Gaussian model is too parsimonious and it overshrinks the school effects η_j towards the global mean. Mixture models are in between these two extremes, although a critical distinction between Dirichlet and NGG models is quite evident from Table 1 and Table 2. Specifically, Dirichlet priors are much closer to the "no pooling" model, displaying a very high number clusters a posteriori, in turn inducing undershrinkage on the school effects. On the other hand, NGG specifications are more conservative, favoring a stronger shrinkage while allowing deviations from Gaussianity whenever needed. In other words, the findings in Table 1 and Table 2 imply that NGG specifications induce the right amount of shrinkage, suitably balancing flexibility and complexity.

We finally investigated the impact of the covariates $\boldsymbol{z}_{ij}$ on the response variable S_{ij}. The posterior distributions of the associated coefficients $\boldsymbol{\gamma}$ are reported in the supplementary material. We found that all the considered covariates explain a considerable fraction of the variability of the response variables. In particular, we noticed a strong positive association between the education level of the parents and the students' performance on the test. Specific employments (e.g. executive positions) were also found to be positively associated with the

School ID	Dirichlet multinomial	NGG multinomial	Gaussian	No pooling
1	0.19	0.11	0.12	0.26
2	0.10	0.08	0.09	0.13
3	0.27	0.15	0.15	0.39
4	0.09	0.05	0.05	0.19
5	-0.02	-0.01	-0.02	-0.04
6	0.68	0.60	0.25	0.90
7	-0.68	-0.69	-0.32	-0.82

Table 2: Posterior mean of the (centred) school effects $\eta_j - \gamma_0$, under different model specifications. Schools with ID from 1 to 5 were randomly chosen from the set of 84 schools. Instead, schools with ID 6 and 7 were identified to better emphasize the over- and under-shrinkage effects.

outcomes. Unsurprisingly, one of the stronger predictors is the regularity of the studies, meaning that students that failed an exam are less likely to obtain a high score at the INVALSI test.

6 Discussion

In this paper we introduce a novel class of discrete random probability measures, termed NIDMs, and provide a complete toolbox for their use as priors for Bayesian inference. They are characterized by a finite number of atoms H and, importantly, homogeneous NRMIs are recovered as a limiting case when $H \to \infty$. Although our results are general, in applications we focus on the NGG special case, which has appealing analytical and computational properties, representing a robust alternative to Dirichlet-multinomial specifications when interest is in the number of clusters K_n. We conclude our presentation by offering a high-level comparison between NIDMs and allied models.

For moderate values of H, there are important differences between NIDMs and the corresponding NRMIs. An illustrative comparison between the Dirichlet-multinomial and the Dirichlet process is given in Green and Richardson (2001), who show that the finite-dimensional specification leads to more "balanced" random partitions. Such a difference carries over to our setting, at least a priori, since the distribution of the weights $(\pi_1, \ldots, \pi_H)$ of a NIDM is

symmetric. Hence, the balancedness of the random partition is a peculiar property of NIDMs which might be desirable or not, depending on the specific application.

Also for large H NIDMs represent a viable alternative to the corresponding homogeneous NRMIs, especially from a computational perspective. Indeed, posterior computations in the latter case usually rely on truncations of the Ferguson and Klass (1972) representation (see e.g. Arbel and Prünster, 2017), which requires numerical approximations while implicitly reducing to a finite-dimensional model. Alternatively, one may use the slice sampling strategy described in Favaro and Teh (2013), which however requires simulating a further layer of latent variables "slicing" the infinite-dimensional model, which could potentially deteriorate the mixing. In contrast, the algorithm we propose for NGG multinomial priors allows to draw iid samples from the posterior distribution of $\tilde{p}_H$ given the data $\theta_1, \ldots, \theta_n$.

Finally, mixture models based on NIDM processes lead to a quite different modeling framework compared to mixture of finite mixtures (MFM) models (Richardson and Green, 1997; Nobile, 2004). In the latter context, a Dirichlet-multinomial process with a prior on H is employed, the inferential target being often, albeit not always, H itself. Instead, our focus is the number of *occupied* clusters $K_n \leq H$, representing the number of *non-empty* mixture components. Although these two quantities are sometimes used interchangeably in applications, they are different objects. This distinction is particularly relevant in species sampling applications, where K_n represents the number of discovered species within the sample, whereas H represents the total number of species in the population. Using NIDMs for species discovery is an interesting research direction. Nonetheless, the MFM framework is a lively research topic, as testified by the recent contributions on asymptotic analysis (De Blasi et al., 2013), on the algorithmical aspects (Miller and Harrison, 2018), and applications to complex data structures (Legramanti et al., 2022). We believe that NIDMs may play a pivotal role also within the context of MFMs, after considering a suitable prior on H. Indeed, we envision that our theoretical findings would enable the derivation of marginal algorithms in the same spirit of Miller and Harrison (2018), while generalizing the current MFM framework. Finally, it is worth noting that while the manuscript was being reviewed, other contributions have appeared proposing alternative modeling approaches that explicitly distinguish occupied clusters and mixture components; see, e.g., Frühwirth-Schnatter et al. (2021), De Blasi and Gyl-Leyva (2021), Nguyen et al. (2021) and Argiento and De Iorio (2022).

References

Arbel, J. and I. Prünster (2017). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing 27*(1), 3–17.

Argiento, R. and M. De Iorio (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *Annals of Statistics*. In press.

Barrios, E., A. Lijoi, L. E. Nieto-Barajas, and I. Prünster (2013). Modeling with Normalized Random Measure Mixture Models. *Statistical Science 28*(3), 313 – 334.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability 31*(4), 929–953.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *The Annals of Statistics 47*(1), 67–92.

Camerlenghi, F., A. Lijoi, and I. Prünster (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics 45*(4), 1062–1091.

Canale, A., A. Lijoi, B. Nipoti, and I. Prünster (2017). On the Pitman–Yor process with spike and slab base measure. *Biometrika 104*(3), 681–697.

Caron, F. and E. B. Fox (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 79*(5), 1295–1366.

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 212–229.

De Blasi, P. and M. F. Gyl-Leyva (2021). Gibbs sampling for mixtures in order of appearance: the ordered allocation sampler. *arXiv:2107.08380*.

De Blasi, P., A. Lijoi, and I. Prünster (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statistica Sinica 23*(3), 1299–1321.

Dunson, D. B., A. H. Herring, and A. M. Siega-Riz (2008). Bayesian inference on changes in response densities over predictor clusters. *Journal of the American Statistical Association 103*(484), 1508–1517.

Durante, D. and D. B. Dunson (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis 13*(1), 29–58.

Durante, D., D. B. Dunson, and J. T. Vogelstein (2017). Nonparametric Bayes modeling of populations of networks. *Journal of the American Statistical Association 112*(520), 1516–1530.

Favaro, S., G. Hadjicharalambous, and I. Prünster (2011). On a class of distributions on the simplex. *Journal of Statistical Planning and Inference 141*(9), 2987–3004.

Favaro, S. and Y. W. Teh (2013). MCMC for normalized random measure mixture models. *Statistical Science 28*(3), 335–359.

Ferguson, T. S. and M. J. Klass (1972). A representation of independent increment processes without Gaussian components. *Annals of Mathematical Statistics 43*(5), 1634—1643.

Frühwirth-Schnatter, S., G. Malsiner-Walli, and B. Grün (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis 16*(4), 1279–1307.

Gelman, A., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing 24*(6), 997–1016.

Gnedin, A. and J. Pitman (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov, POMI 325*, 83–102.

Green, P. J. and S. Richardson (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics 28*(2), 355–375.

Ishwaran, H. and M. Zarepour (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics 30*(2), 269–283.

James, L. F., A. Lijoi, and I. Prünster (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics 33*(1), 105–120.

James, L. F., A. Lijoi, and I. Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics 36*(1), 76–97.

Kallenberg, O. (2017). *Random measures, theory and applications*. Springer.

Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 37*(1), 1–22.

Legramanti, S., T. Rigon, D. Durante, and D. B. Dunson (2022). Extended stochastic block models with application to criminal networks. *Annals of Applied Statistics 16*(4), 2369–2395.

Lijoi, A., R. H. Mena, and I. Prünster (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association 100*(472), 1278–1291.

Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 69*(4), 715–740.

Lijoi, A. and I. Prünster (2010). Models beyond the Dirichlet process. In N. L. Hjort, C. C. Holmes, P. Muller, and S. G. Walker (Eds.), *Bayesian Nonparametrics*. Cambridge University Press.

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing 26*, 303–324.

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356.

Nguyen, T. D., J. Huggins, L. Masoero, L. Mackey, and T. Broderick (2021). Independent finite approximations for Bayesian nonparametric inference. *arXiv:2009.1078v3*.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics 41*(1), 370–400.

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *Annals of Statistics 32*(5), 2044–2073.

Petrone, S., M. Guindani, and A. E. Gelfand (2009). Hybrid Dirichlet mixture models for functional data. *Journal of the American Statistical Association 71*(4), 755–782.

Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme. *Statistics, Probability and Game Theory 30*, 245–267.

Pitman, J. and M. Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability 25*(2), 855–900.

Regazzini, E., A. Lijoi, and I. Prünster (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics 31*(2), 560–585.

Rennie, B. and A. J. Dobson (1969). On Stirling numbers of the second kind. *Journal of Combinatorial Theory 6*, 116–121.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B 59*(4), 768–769.

Ridout, M. S. (2009). Generating random numbers from a distribution specified by its Laplace transform. *Statistics and Computing 19*(4), 439–450.

Rigon, T., D. Durante, and N. Torelli (2019). Bayesian semiparametric modelling of contraceptive behaviour in India via sequential logistic regressions. *Journal of the Royal Statistical Society. Series A: Statistics in Society 182*(1), 225–247.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 73*(5), 689–710.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing 27*(5), 1413–1432.

Xu, Y., P. Müller, and D. Telesca (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics 72*, 955–964.

APPENDIX

A Proofs and additional theoretical results

A.1 Laplace functional of a IDM random measure

The Laplace functional of a IDM random measure is readily obtained after noting that $(\tilde{\mu}_H \mid \tilde{\theta}_1, \ldots, \tilde{\theta}_H)$ is a completely random measure with a purely atomic baseline distribution, placing mass on $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$. Thus given the atoms $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$ for any non-negative function f we get

$$\begin{aligned}
\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_H(\mathrm{d}\theta)}\right) &= \mathbb{E}\left(\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_H(\mathrm{d}\theta)} \mid \tilde{\theta}_1, \ldots, \tilde{\theta}_H\right)\right) \\
&= \mathbb{E}\left(\exp\left\{-\frac{c}{H}\sum_{h=1}^{H}\int_{\mathbb{R}^+}(1-\mathrm{e}^{-sf(\tilde{\theta}_h)})\rho(s)\,\mathrm{d}s\right\}\right) \\
&= \mathbb{E}\left(\exp\left\{-\frac{c}{H}\sum_{h=1}^{H}\psi(f(\tilde{\theta}_h))\right\}\right) \\
&= \prod_{h=1}^{H}\mathbb{E}\left(\exp\left\{\frac{c}{H}\psi(f(\tilde{\theta}_h))\right\}\right) = \left(\int_\Theta \exp\left\{-\frac{c}{H}\psi(f(\theta))\right\}P(\mathrm{d}\theta)\right)^H.
\end{aligned}$$

The last two equalities follows because the locations $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$ are iid from P. The Laplace transform of $\tilde{\mu}_H(A)$ readily follows having set $f = \lambda I_A(x)$, for $\lambda > 0$ and $A \in \Theta$. Indeed, simple calculus lead to

$$\begin{aligned}
\int_\Theta \exp\left\{-\frac{c}{H}\psi(\lambda I_A(\theta))\right\}P(\mathrm{d}\theta) &= \int_A \exp\left\{-\frac{c}{H}\psi(\lambda I_A(\theta))\right\}P(\mathrm{d}\theta) \\
&\quad + \int_{\Theta\setminus A}\exp\left\{-\frac{c}{H}\psi(\lambda I_A(\theta))\right\}P(\mathrm{d}\theta) \\
&= P(A)\exp\left\{-\frac{c}{H}\psi(\lambda)\right\} + 1 - P(A).
\end{aligned}$$

A.2 Moments of a NIDM process

Here we confine ourseleves to considering the first two moments, which have simple analytical forms. Let us recall that $\tilde{p}_{0,H} = (1/H)\sum_{h=1}^{H}\delta_{\tilde{\theta}_h}$, with $\tilde{\theta}_h \overset{\text{iid}}{\sim} P$.

Proposition 1. *Let* $\tilde{p}_H \sim \text{NIDM}(c,\rho;P)$ *and define* $\mathcal{I}(c,\rho) = c\int_{\mathbb{R}^+} u\mathrm{e}^{-c\psi(u)}\tau_2(u)\mathrm{d}u$. *Moreover, let* $A, A_1, A_2 \in \mathscr{B}(\Theta)$ *and set* $C := A_1 \cap A_2$. *Then* $\mathbb{E}(\tilde{p}_H(A)) = P(A)$ *and*

$$\begin{aligned}\mathrm{Var}(\tilde{p}_H(A)) &= P(A)(1-P(A))\left(\mathcal{I}(c,\rho) + \frac{1-\mathcal{I}(c,\rho)}{H}\right),\\ \mathrm{Cov}(\tilde{p}_H(A_1),\tilde{p}_H(A_2)) &= [P(C)-P(A_1)P(A_2)]\left(\mathcal{I}(c,\rho) + \frac{1-\mathcal{I}(c,\rho)}{H}\right).\end{aligned}$$

Unsurprisingly, when $H\to\infty$ the moments of a NIDM process converge to those of a NRMI.

Proof. First, notice that $\mathbb{E}(\tilde{p}_H(A)) = \sum_{h=1}^{H}\mathbb{E}(\pi_h)\mathbb{E}(\delta_{\tilde{\theta}_h}(A)) = P(A)\sum_{h=1}^{H}\mathbb{E}(\pi_h) = P(A)$. As an application of the well-known variance decomposition

$$\mathrm{Var}(\tilde{p}_H(A)) = \mathbb{E}(\mathrm{Var}(\tilde{p}_H(A)\mid\tilde{p}_{0,H})) + \mathrm{Var}(\mathbb{E}(\tilde{p}_H(A)\mid\tilde{p}_{0,H})).$$

Let us focus on the second summand on the right-hand side of the above equation, which is equal to

$$\mathrm{Var}(\mathbb{E}(\tilde{p}_H(A)\mid\tilde{p}_{0,H})) = \mathrm{Var}(\tilde{p}_{0,H}(A)) = \frac{P(A)(1-P(A))}{H}.$$

As for $\mathbb{E}(\mathrm{Var}(\tilde{p}_H(A)\mid\tilde{p}_{0,H}))$, because of Proposition 1 in James et al. (2006) we obtain

$$\begin{aligned}\mathbb{E}(\mathrm{Var}(\tilde{p}_H(A)\mid\tilde{p}_{0,H})) &= \mathbb{E}(\tilde{p}_{0,H}(A)(1-\tilde{p}_{0,H}(A)))\mathcal{I}(c,\rho)\\ &= P(A)(1-P(A))\mathcal{I}(c,\rho) - \mathcal{I}(c,\rho)\mathrm{Var}(\tilde{p}_{0,H}(A))\\ &= P(A)(1-P(A))\left(\mathcal{I}(c,\rho) - \frac{\mathcal{I}(c,\rho)}{H}\right),\end{aligned}$$

from which the result follows. As for the covariance, note that $\mathrm{Var}(\tilde{p}_H(A_1),\tilde{p}_H(A_2)) = P(A)(1-P(A))\mathbb{E}\left(\sum_{h=1}^{H}\pi_h^2\right)$ and $\mathrm{Cov}(\tilde{p}_H(A_1),\tilde{p}_H(A_2)) = (P(C)-P(A_1)P(A_2))\mathbb{E}\left(\sum_{h=1}^{H}\pi_h^2\right)$, meaning that

$$\mathbb{E}\left(\sum_{h=1}^{H}\pi_h^2\right) = \mathcal{I}(c,\rho) + \frac{1-\mathcal{I}(c,\rho)}{H},$$

from which the result follows.

A.3 Proof of Theorem 1

Recall that the Laplace functional can be written as

$$\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_H(\mathrm{d}\theta)}\right) = \mathbb{E}\left(\exp\left\{-\frac{c}{H}\sum_{h=1}^{H}\psi(f(\tilde{\theta}_h))\right\}\right).$$

Now note that the expectations of each $\psi(\tilde{\theta}_h)$ equals $\mathbb{E}(\psi(f(\tilde{\theta}_h))) = \int_\Theta \psi(f(\theta))P(\mathrm{d}\theta) < \infty$, which is finite by assumption. Hence, as an application of the strong law of large numbers, we get

$$\frac{1}{H}\sum_{h=1}^{H}\psi(f(\tilde{\theta}_h)) \xrightarrow{\text{a.s.}} \int_\Theta \psi(f(\theta))P(\mathrm{d}\theta), \qquad H \to \infty,$$

which implies that $\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_H(\mathrm{d}\theta)}\right) \to \mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_\infty(\mathrm{d}\theta)}\right)$ because of bounded convergence theorem.

A.4 Proof of Theorem 2

The symmetry among the weights implies that

$$\Pi_H(n_1,\ldots,n_k) = \frac{H!}{(H-k)!}\mathbb{E}\left(\prod_{j=1}^{k}\pi_j^{n_j}\right).$$

Recalling that $\tilde{\mu}(\Theta) = \sum_{h=1}^{H} J_h$, then we have

$$\begin{aligned}
\mathbb{E}\left(\prod_{j=1}^{k} \pi_j^{n_j}\right) &= \frac{1}{\Gamma(n)} \int_{\mathbb{R}^+} u^{n-1} \mathbb{E}\left(\mathrm{e}^{-u\tilde{\mu}(\Theta)} \prod_{j=1}^{k} J_j^{n_j}\right) \mathrm{d}u \\
&= \frac{1}{\Gamma(n)} \int_{\mathbb{R}^+} u^{n-1} \prod_{j'=k+1}^{H} \mathbb{E}\left(\mathrm{e}^{-uJ_{j'}}\right) \prod_{j=1}^{k} \mathbb{E}\left(\mathrm{e}^{-uJ_j} J_j^{n_j}\right) \mathrm{d}u \\
&= \frac{1}{\Gamma(n)} \int_{\mathbb{R}^+} u^{n-1} \mathrm{e}^{-c\frac{H-k}{H}\psi(u)} \prod_{j=1}^{k} (-1)^{n_j} \frac{\partial^{n_j}}{\partial u^{n_j}} \mathrm{e}^{-\frac{c}{H}\psi(u)} \mathrm{d}u \\
&= \frac{1}{\Gamma(n)} \int_{\mathbb{R}^+} u^{n-1} \mathrm{e}^{-c\psi(u)} \prod_{j=1}^{k} \mathcal{V}_{n_j,H}(u) \mathrm{d}u,
\end{aligned}$$

which concludes the proof, since $\mathcal{V}_{n_j,H}(u) = \frac{c}{H}\Delta_{n_j,H}(u)$. The predictive distributions of Corollary 2 can be obtained exploiting their relationship with the EPPF and after some algebraic manipulation. To obtain the alternative representation (10), recall the following equality, whose proof can be found in Camerlenghi et al. (2019), which holds for $m \geq 1$

$$\mathcal{V}_{m,H}(u) = \frac{c}{H} \sum_{\ell=1}^{m} \xi_{m,\ell,H}(u), \qquad \xi_{m,\ell,H}(u) = \left(\frac{c}{H}\right)^{\ell-1} \frac{1}{\ell!} \sum_{\boldsymbol{q}} \binom{m}{q_1, \ldots, q_\ell} \prod_{r=1}^{\ell} \tau_{q_r}(u), \tag{20}$$

for $\ell = 1, \ldots, m$, where the sum runs over all the vectors of positive integers $\boldsymbol{q} = (q_1, \ldots, q_\ell)$ such that $|\boldsymbol{q}| = m$. Thus, on the light of (20) we can write the EPPF as

$$\begin{aligned}
\Pi_H(n_1, \ldots, n_k) &= \frac{H!}{(H-k)!} \frac{1}{\Gamma(n)} \int_{\mathbb{R}^+} u^{n-1} \mathrm{e}^{-c\psi(u)} \prod_{j=1}^{k} \mathcal{V}_{n_j,H}(u) \mathrm{d}u \\
&= \frac{H!}{(H-k)!} \sum_{\boldsymbol{\ell}} \frac{1}{H^{|\boldsymbol{\ell}|}} \prod_{j=1}^{k} \frac{1}{\ell_j!} \sum_{\boldsymbol{q}_j} \binom{n_j}{q_{j1}, \ldots, q_{j\ell_j}} \Pi_\infty(q_{11}, \ldots, q_{1\ell_1}, \ldots, q_{k1}, \ldots, q_{k\ell_k}),
\end{aligned}$$

where the first sum runs over all vectors $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ such that $\ell_j \in \{1, \ldots, n_j\}$, and the jth of the k sums runs over all the vectors $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{j\ell_j})$ such that $q_{jr} \geq 1$ and $|\boldsymbol{q}_j| = n_j$.

A.5 Proof of Theorem 3

Let us consider the ratio among the two EPPFs, which is equal for any $k \leq H$ to

$$\frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_\infty(n_1,\ldots,n_k)} = \frac{H!}{H^k(H-k)!}\frac{\int_{\mathbb{R}^+} u^{n-1}\mathrm{e}^{-c\psi(u)}\prod_{j=1}^k \Delta_{n_j,H}(u)\mathrm{d}u}{\int_{\mathbb{R}^+} u^{n-1}\mathrm{e}^{-c\psi(u)}\prod_{j=1}^k \tau_{n_j}(u)\mathrm{d}u}.$$

The result follows after noting that the ratio $\frac{H!}{H^k(H-k)!} \leq 1$, and also

$$\frac{\int_{\mathbb{R}^+} u^{n-1}\mathrm{e}^{-c\psi(u)}\prod_{j=1}^k \Delta_{n_j,H}(u)\mathrm{d}u}{\int_{\mathbb{R}^+} u^{n-1}\mathrm{e}^{-c\psi(u)}\prod_{j=1}^k \tau_{n_j}(u)\mathrm{d}u} \geq 1.$$

The latter inequality can be easily obtained from (20), from which is clear that $\Delta_{m,H}(u) = \tau_m(u) + g_m(u)$, where $g_m(u)$ is a positive function, implying that $\Delta_{m,H}(u) \geq \tau_m(u)$ for any $m \geq 1$ and $u > 0$.

A.6 Proof of Theorem 4

The starting point of this proof is based on Corollary 2 in Camerlenghi et al. (2018), from which one can show that

$$\mathbb{P}(K_{n,H} = k) = \sum_{t=k}^{n} \mathbb{P}(K_{n,\infty} = t)\mathbb{P}(K_{t,0} = k), \tag{21}$$

where $K_{n,\infty}$ and $K_{n,H}$ are defined as before, while $K_{n,0}$ for any $n \geq 1$ is the number of distinct values from a sample of n exchangeable observations having prior $\tilde{p}_{0,H}$. The distribution $\mathbb{P}(K_{n,0} = k)$ can be deduced from the associated EPPF, which is

$$\Pi_0(n_1,\ldots,n_k) = \frac{H!}{(H-k)!}H^{-n}, \qquad k \leq H,$$

implying that the distribution of $K_{n,0}$ is given for any $k \leq \min\{H, n\}$

$$\begin{aligned}\mathbb{P}(K_{n,0} = k) &= \frac{1}{k!} \sum_{(n_1,\ldots,n_k)} \binom{n}{n_1, \ldots, n_k} \Pi_0(n_1, \ldots, n_k) \\ &= \frac{H!}{(H-k)!} H^{-n} \frac{1}{k!} \sum_{(n_1,\ldots,n_k)} \binom{n}{n_1, \ldots, n_k} \\ &= \frac{H!}{(H-k)!} H^{-n} \mathscr{S}(n, k).\end{aligned}$$

where the sum runs over all the k-dimensional vectors of positive integers $\boldsymbol{n} = (n_1, \ldots, n_k)$ such that $|\boldsymbol{n}| = n$. The first part of the theorem then follows after the change of variable $\ell = t - k$ in (21). As for the second part, note that the expected value of $K_{n,H}$ can be written as

$$\begin{aligned}\mathbb{E}(K_{n,H}) = \mathbb{E}(\mathbb{E}(K_{n,H} \mid \tilde{\theta}_1, \ldots, \tilde{\theta}_H)) &= \mathbb{E}\left(\mathbb{E}\Big(\sum_{h=1}^{H} I(\tilde{\theta}_h \in \{\theta_1, \ldots, \theta_n\} \mid \tilde{\theta}_1, \ldots, \tilde{\theta}_H\Big)\right) \\ &= \sum_{h=1}^{H} \mathbb{E}(1 - \mathbb{P}(\theta_1 \neq \tilde{\theta}_h, \ldots, \theta_n \neq \tilde{\theta}_h \mid \tilde{\theta}_1, \ldots, \tilde{\theta}_H)) \\ &= \sum_{h=1}^{H} (1 - \mathbb{E}((1-\pi_h)^n)) = H - H\mathbb{E}((1-\pi_1)^n).\end{aligned}$$

The randomness in these equations is given both by $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$ and $\tilde{\theta}_1, \ldots, \tilde{\theta}_H$, whereas in the last step we have used the symmetricity of the weights of a NIDM process. Moreover, with the same steps as for the proof of Theorem 2, one can easily show that

$$\mathbb{E}((1-\pi_1)^n) = \sum_{\ell=1}^{n} \left(1 - \frac{1}{H}\right)^{\ell} \frac{1}{\ell!} \sum_{\boldsymbol{q}} \binom{n}{q_1, \ldots, q_\ell} \Pi_\infty(q_1, \ldots, q_\ell),$$

where the sums runs over $\boldsymbol{q} = (q_1, \ldots, q_\ell)$ such that $q_j \geq 1$ and $|\boldsymbol{q}| = n$, from which the second part of the Theorem follows.

A.7 Bounding the distribution of the number of clusters

The actual evaluation of the probability distribution of $K_{n,H}$ in Theorem 4 might be cumbersome due to the presence of the Stirling numbers. Thus, in cases where it is more convenient to rely on the probability distribution of $K_{n\infty}$ it may be interesting to provide simple bounds for the ratio $\mathbb{P}(K_{n,H}=k)/\mathbb{P}(K_{n,\infty}=k)$. This is achieved in the next Theorem.

Lemma 1. *For any* $k \leq \min\{H, n-1\}$

$$\frac{H!}{H^k(H-k)!} \leq \frac{\mathbb{P}(K_{n,H}=k)}{\mathbb{P}(K_{n,\infty}=k)} \leq \frac{H!}{H^k(H-k)!}\left(1+\frac{1}{2}\sum_{\ell=1}^{n-k}\left(\frac{k}{H}\right)^{\ell}\binom{\ell+k}{k}\frac{\mathbb{P}(K_{n,\infty}=\ell+k)}{\mathbb{P}(K_{n,\infty}=k)}\right),$$

whereas when $k=n=H$*, it holds* $\mathbb{P}(K_{n,H}=n)/\mathbb{P}(K_{n,\infty}=n) = H^{-n}H!/(H-n)!$.

Interestingly, the lower bound in the above Lemma does not depend on the specific NIDM process, and actually coincide with the one obtained by Ishwaran and Zarepour (2002) in the special case of the Dirichlet multinomial NIDM. Instead, the upper bound can be lower than 1, and therefore it is usually tighter than the one already known for the Dirichlet prior. Hence, besides being a generalization to all NIDM processes, Lemma 1 also yields an improvement over existing bounds.

Proof. Recall that the ratio of interest is given by

$$\frac{\mathbb{P}(K_{n,H}=k)}{\mathbb{P}(K_{n,\infty}=k)} = \frac{H!}{H^k(H-k)!}\sum_{\ell=0}^{n-k}\frac{1}{H^{\ell}}\mathscr{S}(\ell+k,k)\frac{\mathbb{P}(K_{n,\infty}=\ell+k)}{\mathbb{P}(K_{n,\infty}=k)},$$

and therefore the lower bound follows naturally. We will write

$$\frac{H!}{H^k(H-k)!} \leq \frac{\mathbb{P}(K_{n,H}=k)}{\mathbb{P}(K_{n,\infty}=k)} = \frac{H!}{H^k(H-k)!}\left(1+\sum_{\ell=1}^{n-k}\frac{1}{H^{\ell}}\mathscr{S}(\ell+k,k)\frac{\mathbb{P}(K_{n,\infty}=\ell+k)}{\mathbb{P}(K_{n,\infty}=k)}\right).$$

Now recall the well-known inequality due to Rennie and Dobson (1969), for which for any $n \geq 2$ and $1 \leq k \leq n-1$ a Stirling number of the second kind can be bounded above by

$$\mathscr{S}(n,k) \leq \frac{1}{2}\binom{n}{k}k^{n-k},$$

implying that we can further bound the summation of the above equation for $1 \leq k \leq \min\{H, n-1\}$ in the following way

$$\sum_{\ell=1}^{n-k} \frac{1}{H^{\ell}} \mathscr{S}(\ell+k,k) \frac{\mathbb{P}(K_{n,\infty}=\ell+k)}{\mathbb{P}(K_{n,\infty}=k)} \leq \frac{1}{2} \sum_{\ell=1}^{n-k} \left(\frac{k}{H}\right)^{\ell} \binom{\ell+k}{k} \frac{\mathbb{P}(K_{n,\infty}=\ell+k)}{\mathbb{P}(K_{n,\infty}=k)}.$$

Hence, the result follows.

A.8 Proof of Theorem 5

We first derive the posterior distribution of $\tilde{p}_{0,H} = (H)^{-1} \sum_{h=1}^{H} \delta_{\tilde{\theta}_h}$ given the $\boldsymbol{\theta}^{(n)}$. This fact is summarized in the following proposition.

Lemma 2. *Let $\theta_1, \ldots, \theta_n$ be a draw from an exchangeable sequence directed by a* NIDM *process. Then, the posterior distribution of $\tilde{p}_{0,H}$ for any $A \in \mathscr{B}(\Theta)$ is*

$$(\tilde{p}_{0,H} \mid \boldsymbol{\theta}^{(n)}) \stackrel{d}{=} \frac{1}{H} \left[\sum_{j=k+1}^{H} \delta_{\bar{\theta}_j} + \sum_{j=1}^{k} \delta_{\theta_j^*} \right],$$

where the atoms $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ are iid draws from P.

Proof. Since the weights of $\tilde{p}_{0,H}$ are fixed and equal, we only need to determine the posterior law of the atoms $(\tilde{\theta}_1, \ldots, \tilde{\theta}_H \mid \boldsymbol{\theta}^{(n)})$. Recall that a NIDM process is a species sampling model, meaning that k out of H atoms are necessarily equal almost surely to one of the previously observed values $\theta_1^*, \ldots, \theta_k^*$, while the remaining $H-k$ are iid draws from the baseline measure P. Notice that the actual order of the weights is irrelevant, because of the symmetry of the weights of $\tilde{p}_{0,H}$. Hence, the result in Lemma 2 follows.

Because of symmetry of the weights, we can assume without loss of generality that the distinct values $\theta_1^*, \ldots, \theta_k^*$ are associated to the first k random weights $\pi_1, \ldots, \pi_k$ of the process $\tilde{p}_H$. The Laplace functional of $\tilde{\mu}_H$, given $\boldsymbol{\theta}^{(n)}$ and $\tilde{p}_{0,H}$ is given by

$$\mathbb{E}\left(\mathrm{e}^{-\tilde{\mu}_H(f)} \mid \boldsymbol{\theta}^{(n)}, \tilde{p}_{0,H}\right) = \frac{\mathbb{E}\left(\mathrm{e}^{-\tilde{\mu}_H(f)} \prod_{j=1}^{k} \pi_j^{n_j} \mid \tilde{p}_{0,H}\right)}{\mathbb{E}\left(\prod_{j=1}^{k} \pi_j^{n_j} \mid \tilde{p}_{0,H}\right)},$$

and hence, with similar steps as for Theorem 2, both at the numerator and the denominator, we obtain

$$
\begin{aligned}
&\mathbb{E}\left(\mathrm{e}^{-\tilde{\mu}_H(f)} \mid \boldsymbol{\theta}^{(n)}, \tilde{p}_{0,H}\right) \\
&= \frac{\int_{\mathbb{R}^+} u^{n-1}\mathrm{e}^{-\frac{c}{H}\sum_{j=1}^{k}\psi(f(\theta_j^*)+u)}\,\mathrm{e}^{-\frac{c}{H}\sum_{h=k+1}^{H}\psi(f(\tilde{\theta}_h)+u)}\prod_{j=1}^{k}\Delta_{n_j,H}(f(\theta_j^*)+u)\mathrm{d}u}{\int_{\mathbb{R}^+} u^{n-1}\mathrm{e}^{-c\psi(u)}\prod_{j=1}^{k}\Delta_{n_j,H}(u)\mathrm{d}u} \\
&= \int_{\mathbb{R}^+} \mathrm{e}^{-\frac{c}{H}\sum_{j=1}^{k}\psi^{(u)}(f(\theta_j^*))}\,\mathrm{e}^{-\frac{c}{H}\sum_{h=k+1}^{H}\psi^{(u)}(f(\tilde{\theta}_h))}\prod_{j=1}^{k}\frac{\Delta_{n_j,H}(f(\theta_j^*)+u)}{\Delta_{n_j,H}(u)}\,q_H(u)\,\mathrm{d}u \\
&= \int_{\mathbb{R}^+} \prod_{h=k+1}^{H}\mathbb{E}\left(\mathrm{e}^{-f(\tilde{\theta}_h)J_h^*}\right)\prod_{j=1}^{k}\mathbb{E}\left(\mathrm{e}^{-f(\theta_j^*)(J_j^*+I_j)}\right)\,q_H(u)\,\mathrm{d}u
\end{aligned}
$$

where we used the fact that $\psi(f(\theta)+u) = \psi^{(u)}(f(\theta)) + \psi(u)$, with $\psi^{(u)}(\lambda)$ denoting the Laplace exponent associated to the tilted jump measure $\rho^*(\mathrm{d}s) = \mathrm{e}^{-us}\rho(s)\mathrm{d}s$. It remains to show that any ratio $\Delta_{m,H}(\lambda+u)/\Delta_{m,H}(u)$ is indeed the Laplace transform associated to some nonnegative random variable, for any $m \geq 1$ and $\lambda > 0$. This is immediately evident from equation (20), because each $\Delta_{m,H}(u)$ can be expressed as a linear combination of Laplace exponents of the form $\tau_m(\lambda+u)/\tau_m(u)$, meaning that each random variable I_j can be interpreted as a mixture of convolutions of random variables. By combining Lemma 2 with the above Laplace functional the result follows.

A.9 Proof of Corollary 4

By exploiting equation (20), one can easily notice that $\mathbb{E}\left(\mathrm{e}^{-\tilde{\mu}_H(f)} \mid \boldsymbol{\theta}^{(n)}, \tilde{p}_{0,H}\right)$ obtained in the proof of Theorem 5 can be interpreted as a mixture over the table configurations. Thus, by augmenting and subsequently conditioning on the table frequencies, one can easily obtain

$$
\begin{aligned}
\mathbb{E}\left(\mathrm{e}^{-\tilde{\mu}_H(f)} \mid \boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}, \tilde{p}_{0,H}\right) &= \int_{\mathbb{R}^+} \mathrm{e}^{-\frac{c}{H}\sum_{h=k+1}^{H}\psi^{(u)}(f(\tilde{\theta}_h))}\,\mathrm{e}^{-\frac{c}{H}\sum_{j=1}^{k}\psi^{(u)}(f(\theta_j^*))} \\
&\quad \times \prod_{j=1}^{k}\prod_{r=1}^{\ell_j}\frac{\tau_{q_{jr}}(f(\theta_j^*)+u)}{\tau_{q_{jr}}(u)} q_\infty(u)\mathrm{d}u \\
&= \int_{\mathbb{R}^+} \prod_{h=k+1}^{H}\mathbb{E}\left(\mathrm{e}^{-f(\tilde{\theta}_h)J_h^*}\right)\prod_{j=1}^{k}\prod_{r=1}^{\ell_j}\mathbb{E}\left(\mathrm{e}^{-f(\theta_j^*)(J_j^*+I_{jr})}\right) q_\infty(u)\mathrm{d}u,
\end{aligned}
$$

from which the result follows, by combining the above equation with Lemma 2.

A.10 Dirichlet multinomial process

In order to derive the EPPF of the Dirichlet multinomial from Theorem 2 one just need to notice that when $\rho(s)\mathrm{d}s = s^{-1}\mathrm{e}^{-s}\mathrm{d}s$, then for any $m \geq 1$ and $u > 0$ it holds

$$\mathcal{V}_{m,H}(u) = \frac{c}{H}\Delta_{m,H}(u) = \frac{\Gamma(m + c/H)}{\Gamma(m)\Gamma(c/H)}\tau_m(u),$$

which can be verified directly from the definition of $\mathcal{V}_{m,H}(u)$ and $\tau_m(u)$. Substituting the above quantity in general formula of Theorem 2, one has simply that for $k \leq H$

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!}\frac{1}{c^k\Gamma(c/H)^k}\prod_{j=1}^{k}\left(\frac{\Gamma(n_j + c/H)}{\Gamma(n_j)}\right) \times \Pi_\infty(n_1, \ldots, n_k),$$

where $\Pi_\infty(n_1, \ldots, n_k)$, the EPPF a Dirichlet process, is $\Pi_\infty(n_1, \ldots, n_k) = c^k \prod_{j=1}^{k}\Gamma(n_j)/(c)_n$. Hence the desired EPPF can be obtained with some simple algebra. Notice that one could also obtain this result specializing the general EPPF of the NGG multinomial process, by letting $\sigma \to 0$. Indeed, recall that in the Dirichlet process case $\mathscr{V}_{n,k} = c^k/(c)_n$, and that as $\sigma \to 0$ one has $\sigma^{-k}\mathscr{C}(n, k; \sigma) \to |s(n, k)|$, the sign-less Stirling number of the first kind. The distribution of $K_{n,H}$ is also obtained by exploiting properties of Stirling numbers. Indeed, specializing Theorem 4 and after a change of variable

$$\begin{aligned}
\mathbb{P}(K_{n,H} = k) &= \frac{H!}{(H-k)!}\frac{1}{(c)_n}\sum_{t=k}^{n}\left(\frac{c}{H}\right)^t \mathscr{S}(t, k)|s(n, t)| \\
&= \frac{H!}{(H-k)!}\frac{(-1)^n}{(c)_n}\sum_{t=k}^{n}\left(-\frac{c}{H}\right)^t \mathscr{S}(t, k)s(n, t) \\
&= \frac{H!}{(H-k)!}\frac{(-1)^k}{(c)_n}\mathscr{C}(n, k; -c/H).
\end{aligned}$$

A.11 NGG multinomial process

Substituting the EPPF of a generalized gamma NRMI in (11), and focusing on the summation one has

$$\frac{1}{\ell_j!}\sum_{\boldsymbol{q}_j}\binom{n_j}{q_{j1},\ldots,q_{j\ell_j}}\Pi_\infty(q_{11},\ldots,q_{1\ell_1},\ldots,q_{k1},\ldots,q_{k\ell_k}) =$$

$$= \mathscr{V}_{n,|\boldsymbol{\ell}|}\frac{1}{\ell_j!}\sum_{\boldsymbol{q}_j}\binom{n_j}{q_{j1},\ldots,q_{j\ell_j}}\prod_{r=1}^{\ell_j}(1-\sigma)_{q_{jr}-1} = \mathscr{V}_{n,|\boldsymbol{\ell}|}\frac{\mathscr{C}(n_j,\ell_j;\sigma)}{\sigma^{\ell_j}},$$

from which the EPPF of a NGG multinomial process follows. With the same reasoning, one also obtain the explicit relation for $\Delta_{m,H}(u)$ after recalling (20).

B Algorithms and computational strategies

B.1 Simulation of U in the NGG multinomial case

We devise here a simple acceptance-rejection method for sampling the latent variable U in the NGG multinomial case, whose density was denoted with $q_H(u)$. Let us focus on the limiting case $H \to \infty$, and suppose we want to simulate a random variable having density proportional to

$$q_\infty(u) \propto u^{n-1}(\kappa+u)^{-n+k\sigma}\exp\left\{-\frac{c}{\sigma}\left[(\kappa+u)^\sigma-\kappa^\sigma\right]\right\}.$$

Rather than handling U directly it is convenient to draw samples from $V := \log U$, whose density function is readily available after a change of variable:

$$q_\infty(v) \propto \mathrm{e}^{vn}(\kappa+\mathrm{e}^v)^{-n+k\sigma}\exp\left\{-\frac{c}{\sigma}\left[(\kappa+\mathrm{e}^v)^\sigma-\kappa^\sigma\right]\right\}.$$

The distribution of V is log-concave, that is, the logarithm of its density is concave, as one can readily verify through direct calculation of the second derivative. This is a major computational advantage and it implies, for instance, that the distribution of V is unimodal. Moreover, we note that everal black-box techniques were developed for sampling log-concave distributions.

We propose a simple sampling algorithm which has the advantage of being straightforward

to implement, and it can be easily extended to the finite-dimensional setting. As a matter of fact, it is just an application of the well-known ratio-of-uniform method, which we recall here for convenience. Set

$$b = \sqrt{\sup\{q_\infty(v) : v \in \mathbb{R}\}},\ b_- = -\sqrt{\sup\{v^2 q_\infty(v) : v \leq 0\}},\ b_+ = \sqrt{\sup\{v^2 q_\infty(v) : v \geq 0\}}.$$

Log-concavity of V ensures that the above constants are finite. Unfortunately, there are no closed form expressions for b, b_- and b_+, but they can be readily computed via univariate numerical maximization, which is a particularly simple problem in this log-concave setting. Then, a draw from U can be obtained as follows:

Step 1. Sample independently $\mathcal{I}_1, \mathcal{I}_2$ uniformly on $(0, b)$ and (b_-, b_+), respectively.

Step 2. Set the candidate value $V^* = \mathcal{I}_2/\mathcal{I}_1$.

Step 3. If $\mathcal{I}_1^2 \leq q_\infty(V^*)$ then accept V^* and set $V = V^*$, otherwise repeat the whole procedure.

Step 4. Set $U = \exp V$.

The simulation from $q_H(u)$ proceeds in a similar manner, with the obvious modifications. A good degree of tractability is preserved because $q_H(u)$, and equivalently $q_H(v)$, is a finite mixture of densities having the kernel of $q_\infty(u)$, namely

$$\begin{aligned} q_H(u) &\propto \sum_{\boldsymbol{\ell}} \left[\prod_{j=1}^{k} \left(\frac{c}{H}\right)^{\ell_j - 1} \frac{\mathscr{C}(n_j, \ell_j; \sigma)}{\sigma^{\ell_j}} \right] \times u^{n-1}(\kappa + u)^{-n+|\boldsymbol{\ell}|\sigma} \exp\left\{ -\frac{c}{\sigma}\left[(\kappa+u)^\sigma - \kappa^\sigma\right] \right\}, \\ &\propto u^{n-1} \exp\left\{ -\frac{c}{\sigma}\left[(\kappa+u)^\sigma - \kappa^\sigma\right] \right\} \prod_{j=1}^{k} \sum_{\ell_j=1}^{n_j} \xi_{n_j,\ell_j,H}(u), \end{aligned}$$

which implies that the constants b, b_- and b_+ involved in the simulation of $q_H(v)$ are finite also in this case. Moreover, as $H \to \infty$ the density $q_H(v)$ converges to $q_\infty(v)$, implying that log-concavity is recovered at the limit.

B.2 Gibbs sampling algorithm for the INVALSI application

We describe here a Gibbs sampling algorithm for posterior computation of the model described in Section 5. Let $G_j \in \{1, \ldots, H\}$ be an indicator function denoting to which mixture component each school is allocated, for $j = 1, \ldots, 84$. The Gibbs sampling algorithm alter-

nates between the following full conditional steps:

Step 1. Exploiting standard results of Gaussian linear models, the full conditional for the coefficients $\boldsymbol{\gamma}$ is multivariate Gaussian with

$$(\boldsymbol{\gamma} \mid -) \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \qquad \tilde{\boldsymbol{\Sigma}} = (\boldsymbol{Z}^\intercal \boldsymbol{Z}/\sigma^2 + \boldsymbol{B}^{-1})^{-1}, \qquad \tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\Sigma}}(\boldsymbol{Z}^\intercal \boldsymbol{\eta}_{\boldsymbol{\gamma}}/\sigma^2 + \boldsymbol{B}^{-1}\boldsymbol{b}),$$

where $\boldsymbol{\eta}_{\boldsymbol{\gamma}}$ is a vector with entries $\eta_{ij\boldsymbol{\gamma}} = S_{ij} - \mu_j$, for $i = 1, \ldots, N_j$ and $j = 1, \ldots, 84$, whereas $\boldsymbol{Z}$ is the corresponding design matrix having row entries $\boldsymbol{z}_{ij}^\intercal$.

Step 2. The full conditional for the residual variance is

$$(\sigma_\epsilon^{-2} \mid -) \sim \text{GAMMA}\left(a_\sigma + N/2, b_\sigma + \frac{1}{2}\sum_{j=1}^{84}\sum_{i=1}^{N_j}(S_{ij} - \mu_j - \boldsymbol{z}_{ij}^\intercal \boldsymbol{\gamma})^2\right),$$

which can be obtained through standard calculations involved in Gaussian linear models.

Step 3. We update the cluster indicators $G_j \in \{1, \ldots, H\}$ from their full conditional categorical random variables

$$\mathbb{P}(G_j = h) = \frac{\pi_h \mathcal{N}(\mu_j; \bar{\mu}_h, \bar{\sigma}_h^2)}{\sum_{h'=1}^{H} \pi_{h'} \mathcal{N}(\mu_j; \bar{\mu}_{h'}, \bar{\sigma}_{h'}^2)}, \qquad h = 1, \ldots, H,$$

for any $j = 1, \ldots, 84$.

Step 4. The full conditional for the school-specific parameters, given the above cluster assignments, is easily available as

$$(\mu_j \mid -) \overset{\text{ind}}{\sim} N\left(\frac{\sum_{i=1}^{N_j}(S_{ij} - \boldsymbol{z}_{ij}^\intercal \boldsymbol{\gamma})/\sigma^2 + \bar{\mu}_{G_j}/\bar{\sigma}_{G_j}^2}{1/\bar{\sigma}_{G_j}^2 + N_j/\sigma^2}, \frac{1}{1/\bar{\sigma}_{G_j}^2 + N_j/\sigma^2}\right),$$

independently for every $j = 1, \ldots, 84$.

Step 5. The full conditional for $\bar{\mu}_h$ and $\bar{\sigma}_h^2$ are given by

$$(\bar{\mu}_h \mid -) \overset{\text{ind}}{\sim} N\left(\frac{\sum_{j:G_j=h} \mu_j/\bar{\sigma}_h^2}{1/\sigma_{\bar{\mu}}^2 + 1/\bar{\sigma}_h^2 \sum_{j=1}^{84} I(G_j = h)}, \frac{1}{1/\sigma_{\bar{\mu}}^2 + 1/\bar{\sigma}_h^2 \sum_{j=1}^{84} I(G_j = h)}\right),$$

independently for $h = 1, \ldots, H$ and

$$(\bar{\sigma}_h^{-2} \mid -) \overset{\text{ind}}{\sim} \text{GAMMA}\left(a_{\bar{\sigma}} + \frac{1}{2}\sum_{j=1}^{84} I(G_j = h), b_{\bar{\sigma}} + \frac{1}{2}\sum_{j:G_j=h} (\mu_j - \bar{\mu}_{G_j})^2\right),$$

again independently for $h = 1, \ldots, H$.

Step 6. Update the weights $(\pi_1, \ldots, \pi_H)$ from their full conditional distribution by exploiting the posterior characterization of Theorem 5 and the simplifications described in Example 6. The frequencies $n_1, \ldots, n_k$ in the notation of Theorem 5 corresponds to the *non-zero elements* of the vector

$$(\bar{n}_1, \ldots, \bar{n}_H) = \left(\sum_{j=1}^{84} I(G_j = 1), \ldots, \sum_{j=1}^{84} I(G_j = H)\right),$$

in any arbitrary order. In first place, we sample the latent random variable U given the frequencies $n_1, \ldots, n_k$ from $q_H(u)$, following the procedure described in Section B.1. Conditionally on U, then one samples the iid random variables $J_1^*, \ldots, J_H^*$ according to a tempered stable distribution (Ridout, 2009), whose parameters are described in Example 6. Finally, conditionally on U, we need to sample the collection of independent random variables $\bar{I}_1, \ldots, \bar{I}_H$ with associated frequencies $\bar{n}_1, \ldots, \bar{n}_H$. For any $h = 1, \ldots, H$ the density of a random variable $\bar{I}_h$ is described in equation (15) of the manuscript when the corresponding frequency $\bar{n}_h > 0$. Samples from (15) can be easily drawn, being a finite mixture of gamma densities. Instead, for any $h = 1, \ldots, H$ such that $\bar{n}_h = 0$ we set $\bar{I}_h = 0$. Hence, a sample from the full conditional is obtained by letting

$$(\pi_1, \ldots, \pi_H) = \left(\frac{J_1^* + \bar{I}_1}{\sum_{h=1}^H (J_h^* + \bar{I}_h)}, \ldots, \frac{J_H^* + \bar{I}_H}{\sum_{h=1}^H (J_h^* + \bar{I}_h)}\right).$$

C Additional material for the INVALSI application

In this Section we provide additional results and plots for the INVALSI application. In Figure 4, we display the posterior distribution of the η_j random effects for 40 randomly selected schools. It is, then, apparent that a certain degree of variability among schools is present and a posterior summary as the median of each η_j might be employed, for example, to identify

virtuous schools. In Figure 4 we also depict the posterior mean of $\sum_{h=1}^{H} \pi_h \mathcal{N}(\bar{\mu}_h, \bar{\sigma}_h^2)$ that stands as an estimate of the data generating density, under three model specifications.

Moreover, in Figures 5-7 we report the posterior distributions of the $\boldsymbol{\gamma}$ coefficients using *violin plots*. Recall that student-specific categorical variables are: the gender of the student, the education level of her/his father and mother (primary school, secondary school, etc.), the employment status of her/his father and mother, the regularity of the student (i.e. regular, in late, etc.), and the citizenship (Italian, first generation immigrant, etc.). To avoid collinearity, the first category is omitted and regarded as baseline.

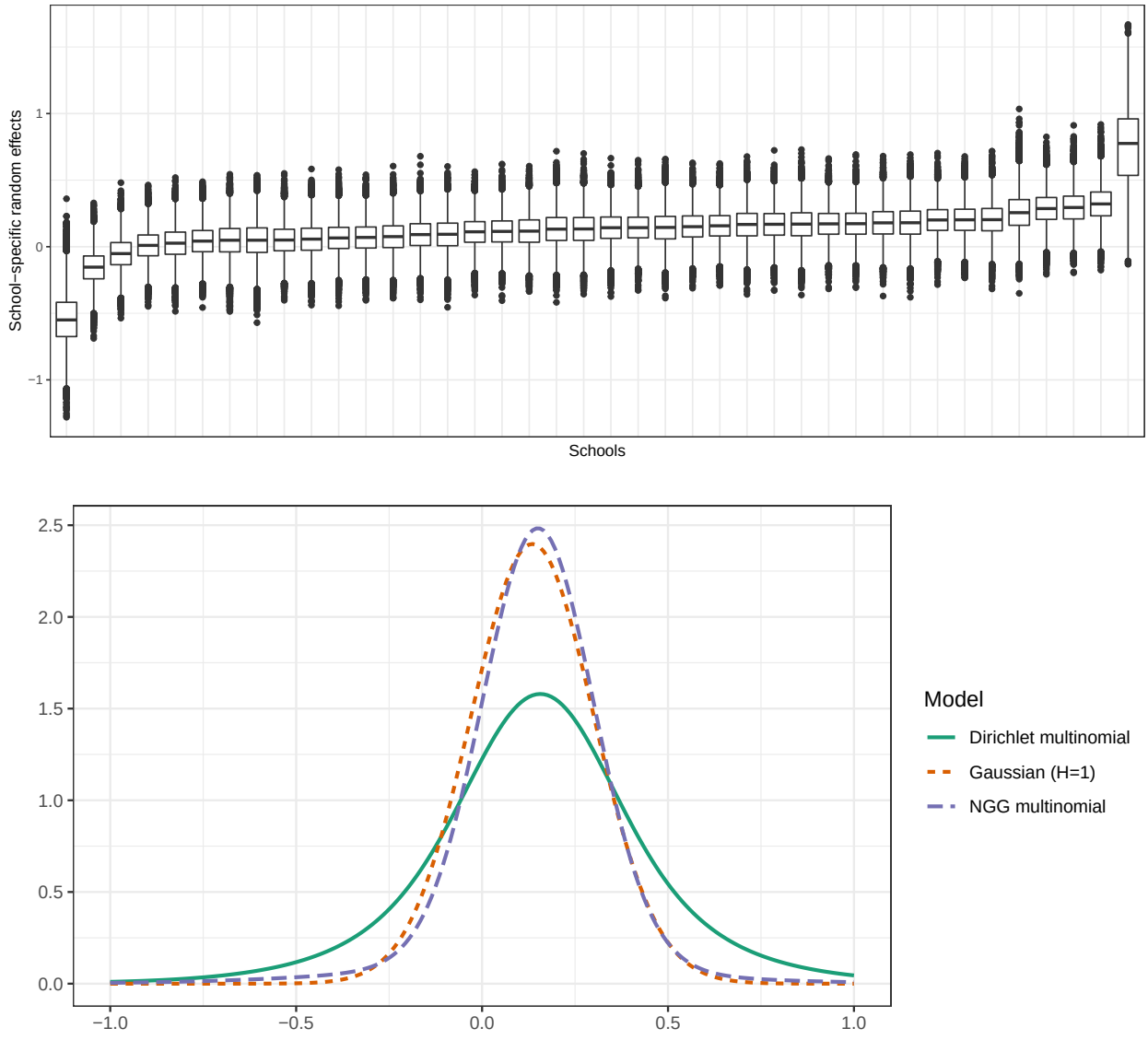

Figure 4: Top panel: Posterior distribution of the random effects η_j for 40 randomly selected schools, ordered according to the posterior median. Bottom panel: Posterior mean of the random density $\sum_{h=1}^{H} \pi_h \mathcal{N}(\bar{\mu}_h, \bar{\sigma}_h^2)$.

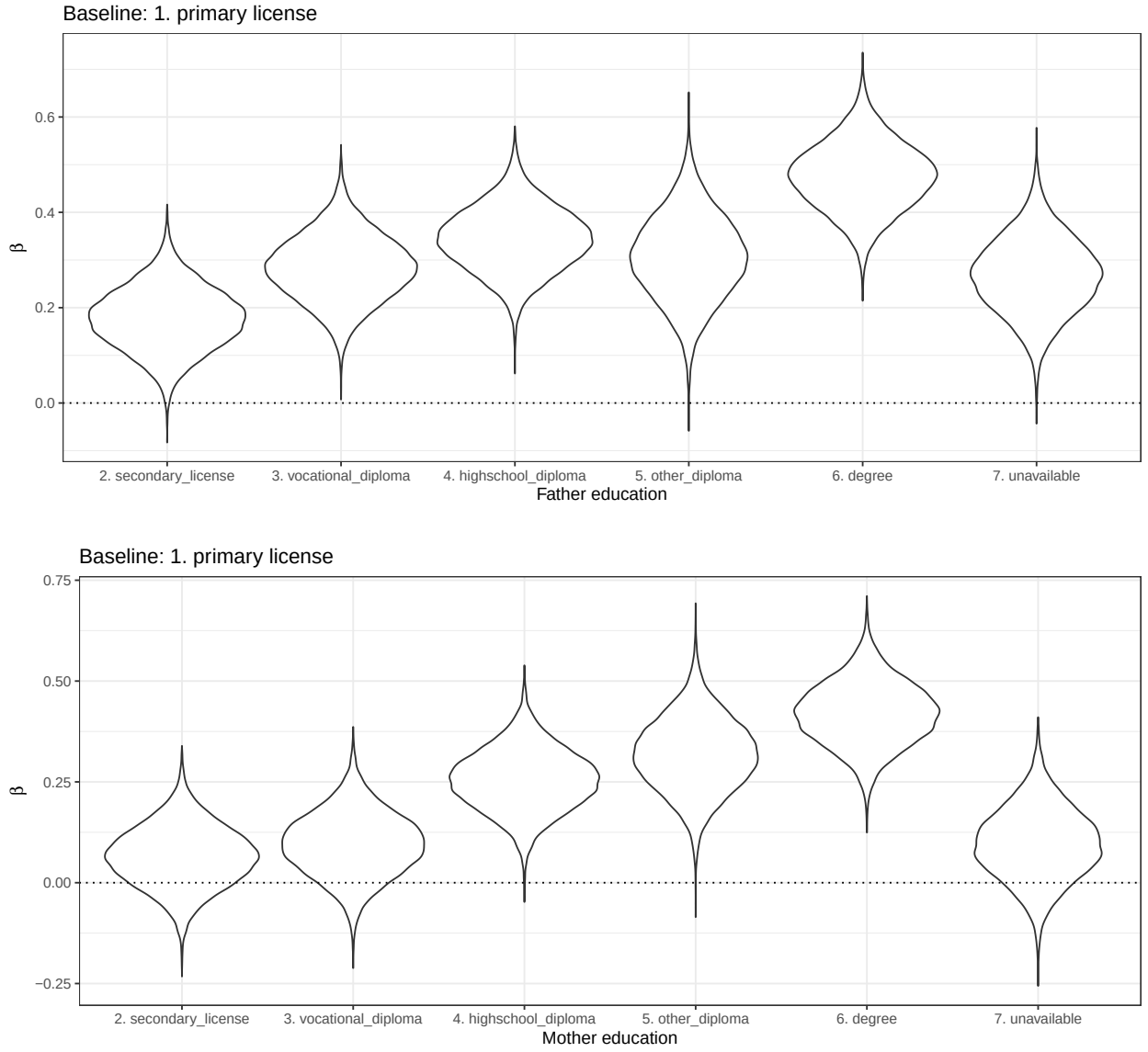

Figure 5: Posterior distribution of the $\boldsymbol{\gamma}$ coefficients in the INVALSI application.

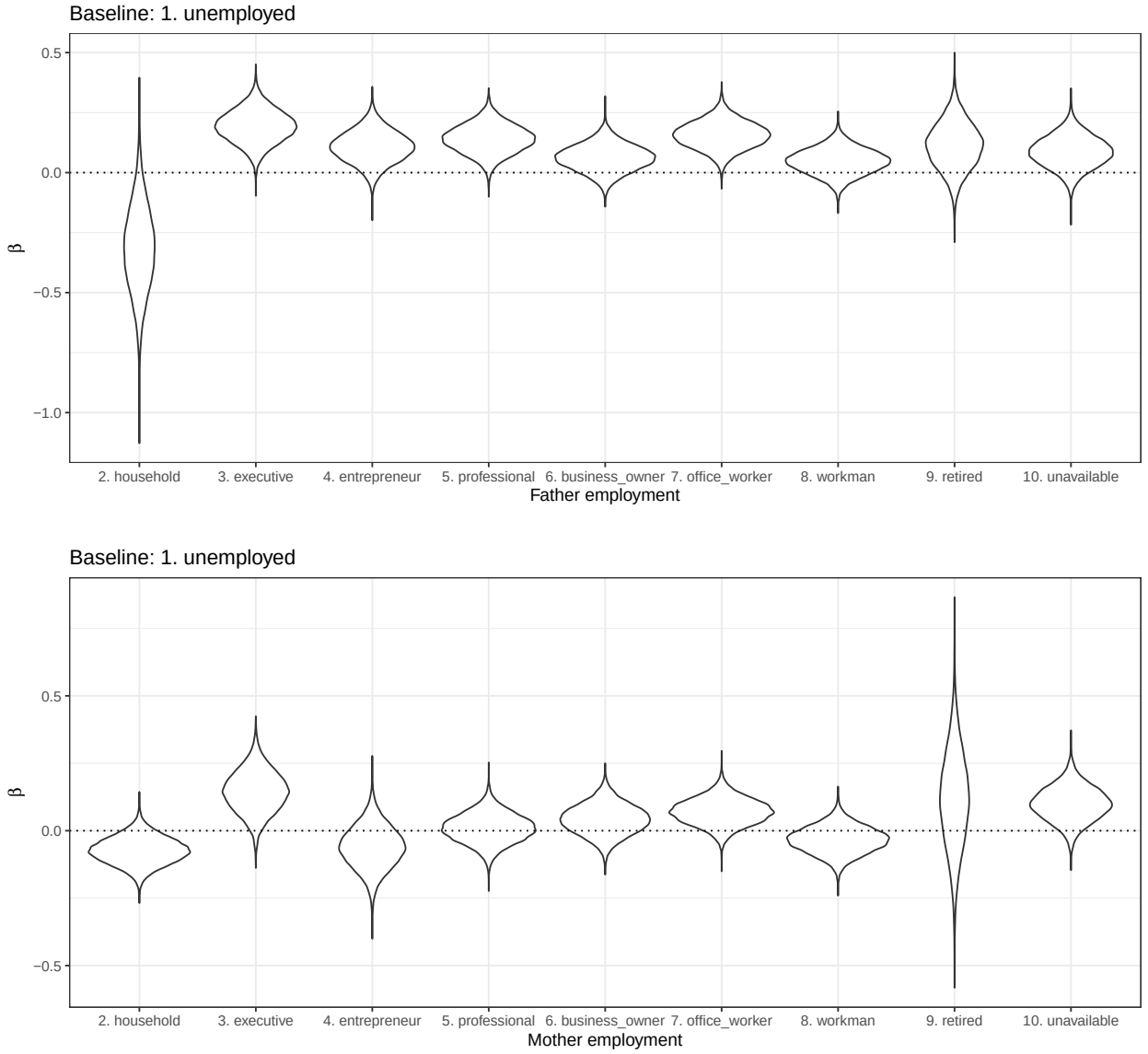

Figure 6: Posterior distribution of the $\boldsymbol{\gamma}$ coefficients in the INVALSI application.

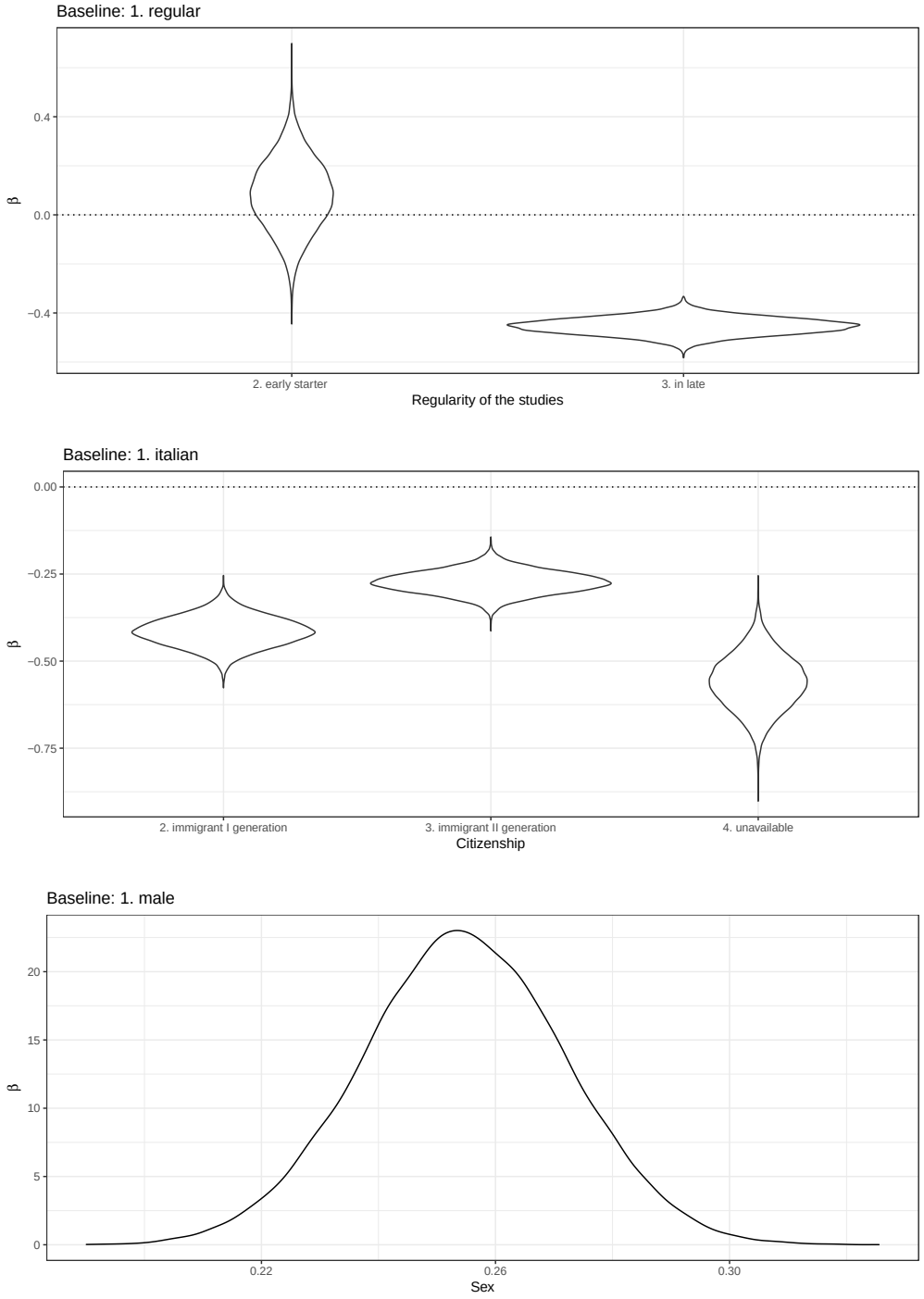

Figure 7: Posterior distribution of the $\boldsymbol{\gamma}$ coefficients in the INVALSI application.

D Overfitted mixture illustration

We empirically illustrate the additional flexibility provided by the NGG multinomial prior on a synthetic dataset. Let us consider a mixture model having density

$$\sum_{h=1}^{H} \pi_h \mathcal{N}(\bar{\mu}_h, \bar{\sigma}_h^2),$$

for a set of observations $Y_1, \ldots, Y_n$. We propose the use of general NIDM processes as mixing measure, corresponding to the following prior specification

$$(\pi_1, \ldots, \pi_{H-1}) \sim \text{NID}\left(\frac{c}{H}, \ldots, \frac{c}{H}; \rho\right), \qquad (\bar{\mu}_h, \bar{\sigma}_h^2) \overset{\text{iid}}{\sim} P, \qquad h = 1, \ldots, H,$$

where P is a diffuse probability measure on $\mathbb{R} \times \mathbb{R}^+$. In our simulation studies, we employ a NGG multinomial process in the above specification, which will be be compared to the Dirichlet multinomial process, the Dirichlet process (DP), and the infinite-dimensional NGG process, in a broad variety of scenarios, hyperparameter settings, and sample sizes. In all these cases, the aim is to infer the true number of mixture components from the data by relying on the overfitted mixture approach (Rousseau and Mengersen, 2011). We also obtain a posterior estimate for the random density $\sum_{h=1}^{H} \pi_h \mathcal{N}(\bar{\mu}_h, \bar{\sigma}_h^2)$. We consider 5 different data generating processes, corresponding to mixture models with different characteristics. Moreover, the analyses are replicated by varying the hyperparameter settings and the sample sizes, for a total of 5 datasets $\times$ 4 scenarios (i.e. hyperparameter settings and sample sizes) $=$ 20 comparisons among the aforementioned DP, DP multinomial, NGG, NGG multinomial models.

The first synthetic dataset (DATASET 1) consists of an independent sample $Y_1, \ldots, Y_n$ of observations from the Gaussian distribution

$$\mathcal{N}(y; 0, 0.25^2),$$

where $\mathcal{N}(y; \bar{\mu}, \bar{\sigma}^2)$ denotes the density function of a Gaussian. In DATASET 1 we aim at showing that a mixture model is able to recover a simple parametric specification. In the second synthetic dataset (DATASET 2), the simulated data $Y_1, \ldots, Y_n$ are independent samples

from the following mixture

$$\frac{1}{2}\mathcal{N}(y; 0, 0.25^2) + \frac{1}{2}\mathcal{N}(y; 1, 0.25^2),$$

corresponding to an equally weighted mixture of Gaussians models with two components. The third synthetic dataset (DATASET 3), is obtained by sampling from the following mixture model

$$\frac{1}{5}\mathcal{N}(y; -2, 0.2^2) + \frac{1}{5}\mathcal{N}(y; -1, 0.2^2) + \frac{1}{5}\mathcal{N}(y; 0, 0.2^2), + \frac{1}{5}\mathcal{N}(y; 1, 0.2^2) + \frac{1}{5}\mathcal{N}(y; 2, 0.2^2).$$

In DATASET 3 the true number of mixture components is 5, and the mixing weights are equal. The fourth synthetic dataset (DATASET 4) is a variant of the second, having considered unbalanced weights. Namely, in DATASET 4 the observations are generated from

$$\frac{1}{5}\mathcal{N}(y; 0, 0.25^2) + \frac{4}{5}\mathcal{N}(y; 1, 0.25^2).$$

In a similar fashion, the fifth and last syntetic dataset (DATASET 5) that we consider is a variant of the third. Indeed, in DATASET 5 the observations are drawn from

$$\frac{1}{15}\mathcal{N}(y; -2, 0.2^2) + \frac{2}{15}\mathcal{N}(y; -1, 0.2^2) + \frac{3}{15}\mathcal{N}(y; 0, 0.2^2), + \frac{4}{15}\mathcal{N}(y; 1, 0.2^2) + \frac{5}{15}\mathcal{N}(y; 2, 0.2^2),$$

thus displaying unequal weights.

Throughout the simulation studies, we assume the conditionally conjugate prior $\bar{\mu}_h \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\bar{\mu}})$, $\bar{\sigma}_h^{-2} \overset{\text{iid}}{\sim} \text{GAMMA}(a_{\bar{\sigma}}, b_{\bar{\sigma}})$, for $h = 1, \ldots, H$, where we set $\sigma^2_{\bar{\mu}} = 1000$ and $a_{\bar{\sigma}} = 2.5$, $b_{\bar{\sigma}} = 0.1$. Moreover, we let $H = 30$, a fairly conservative upper bound for the true number of mixture components. The choice of the other hyperparameters will depend on the specific scenario, as detailed in the subsequent Sections. Posterior samples for all the relevant quantities, such as $K_{n,H}$, can be obtained via MCMC through Gibbs sampling steps similar to those described in Section B.2. We run the algorithms for 12′000 iterations and hold out the first 2′000 as burn-in period.

D.1 Scenarios of the simulation study

We consider a total of 4 scenarios (A, B, C, D), each with a specific sample size: $n = 100, 300, 600$ and then again $n = 100$. For all the 5 datasets we consider a sample $Y_1, \ldots, Y_n$, with n being determined by the corresponding scenario. In each scenario, we also consider four different prior specifications having roughly the same a priori expected value $\mathbb{E}(K_{n,H})$, for suitable sets of parameters c, κ, σ, and H that correspond to the Dirichlet multinomial process, the Dirichlet process (DP), and the infinite-dimensional NGG process. The only exception is represented by the last scenario (SCENARIO D), in which the expected value $\mathbb{E}(K_{n,H})$ is different across the a priori specifications. The details of the parameter settings are reported in Table 3, Table 4, Table 5, and Table 6.

Table 3: Hyperparameter settings for the simulation study in SCENARIO A, $n = 100$

Hyperparameters	Dirichlet multinomial	DP	NGG multinomial	NGG
c	21.9	8.2	0.1	0.22
κ	1	1	0.2	1
σ	0	0	0.8	0.6
H	30	∞	30	∞
$E(K_{n,H})$	21.6	21.6	21.6	21.6

Table 4: Hyperparameter settings for the simulation study in SCENARIO B, $n = 300$

Hyperparameters	Dirichlet multinomial	DP	NGG multinomial	NGG
c	17.4	6.16	0.1	0.18
κ	1	1	1	1
σ	0	0	0.7	0.5
H	30	∞	30	∞
$E(K_{n,H})$	24.6	24.6	24.6	24.6

D.2 Summary of the results

The results of the simulation studies are reported in Figures 8–27, displaying the a priori and the a posteriori distribution of the number of clusters, together with the posterior mean

Table 5: Hyperparameter settings for the simulation study in SCENARIO C, $n = 600$

Hyperparameters	Dirichlet multinomial	DP	NGG multinomial	NGG
c	10.6	4.6	0.2	0.16
κ	1	1	0.2	0.2
σ	0	0	0.6	0.45
H	30	∞	30	∞
$E(K_{n,H})$	23.0	23.0	23.0	23.0

Table 6: Hyperparameter settings for the simulation study in SCENARIO D, $n = 100$

Hyperparameters	Dirichlet multinomial	DP	NGG multinomial	NGG
c	0.60	0.55	0.1	0.22
κ	1	1	0.2	1
σ	0	0	0.8	0.6
H	30	∞	30	∞
$E(K_{n,H})$	3.5	3.5	21.6	21.6

of the density.

In all scenarios and datasets, it is apparent that, under the Dirichlet multinomial and Dirichlet process specifications, the distribution of $K_{n,H}$ struggles to deviate from the prior. Conversely, for both the NGG and the NGG multinomial cases, the posterior law of $K_{n,H}$ correctly recovers the true number of mixture components even when the mean of the prior distribution is far from the true number. This behavior motivates the use of the NGG multinomial to robustify mixture modeling. The estimates of the NGG and NGG multinomial are quite similar, as one would expect given the theoretical findings that are illustrated in the main paper. As expected, the effect becomes somewhat less pronounced when the sample size is high (SCENARIO C), because there is more information in the data. Conversely, when the sample size is low (SCENARIO A) and the mixtures are not perfectly separated (e.g. Figure 10) the differences are important. In SCENARIO D we compare the case in which Dirichlet-based priors are well-calibrated, whereas NGG priors are not. Insterestingly, despite this disadvantage, the NGG priors lead to a posterior distribution for $K_{n,H}$ that is similar to that of the well-calibrated Dirichlet specifications.

In terms of density estimation, the four procedures lead to comparable results when

the sample size is sufficiently high, namely when $n = 600$. Conversely, when the sample size is low, i.e. when $n = 100$, a misscalibrated prior choice for $K_{n,H}$ has an impact also the posterior estimate for the density. For instance, in Figure 8 the Dirichlet multinomial specification has heavier tails than the NGG, since it is capturing the spikes present in the tails of the distribution, although the latter can be regarded as noise.

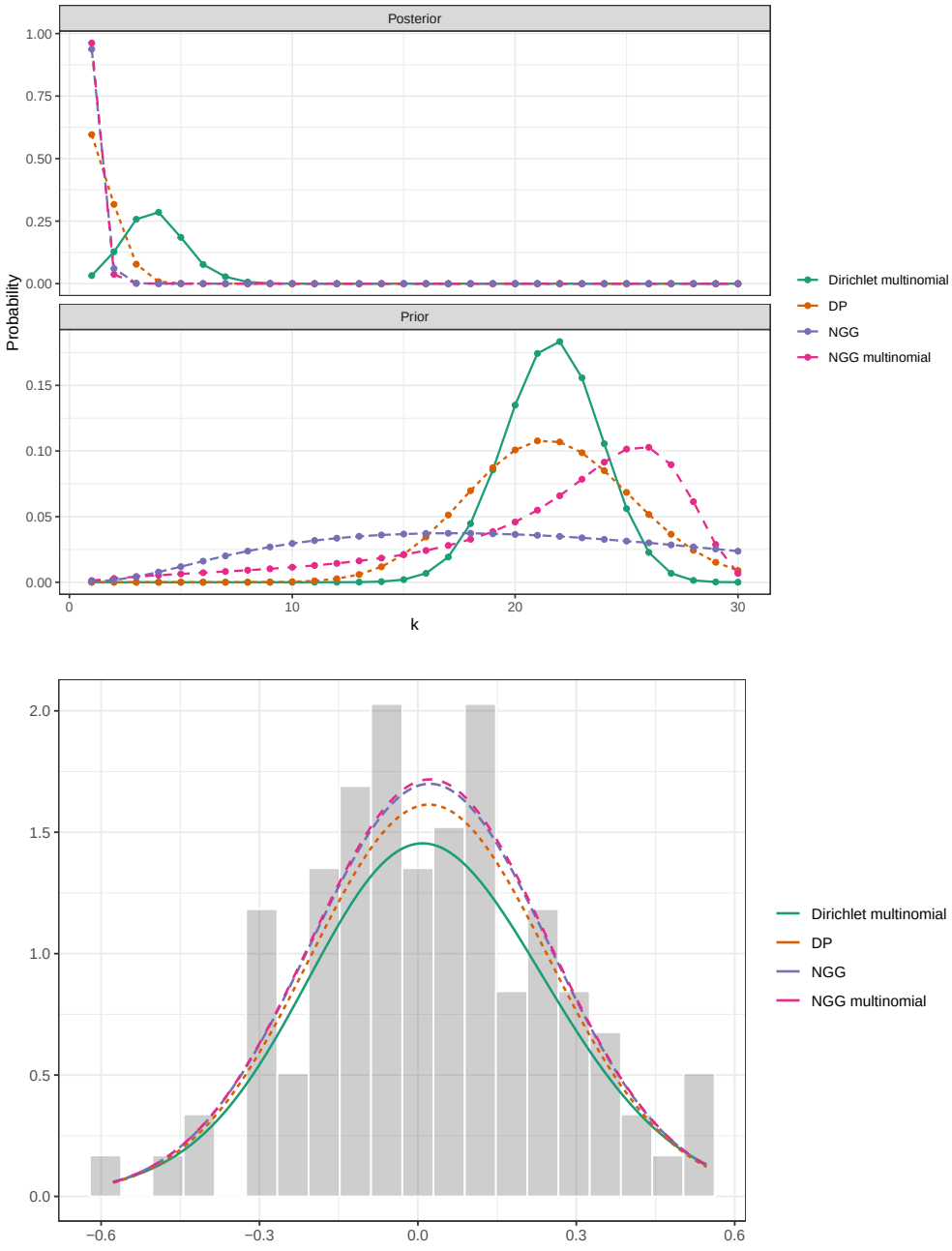

Figure 8: DATASET 1, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

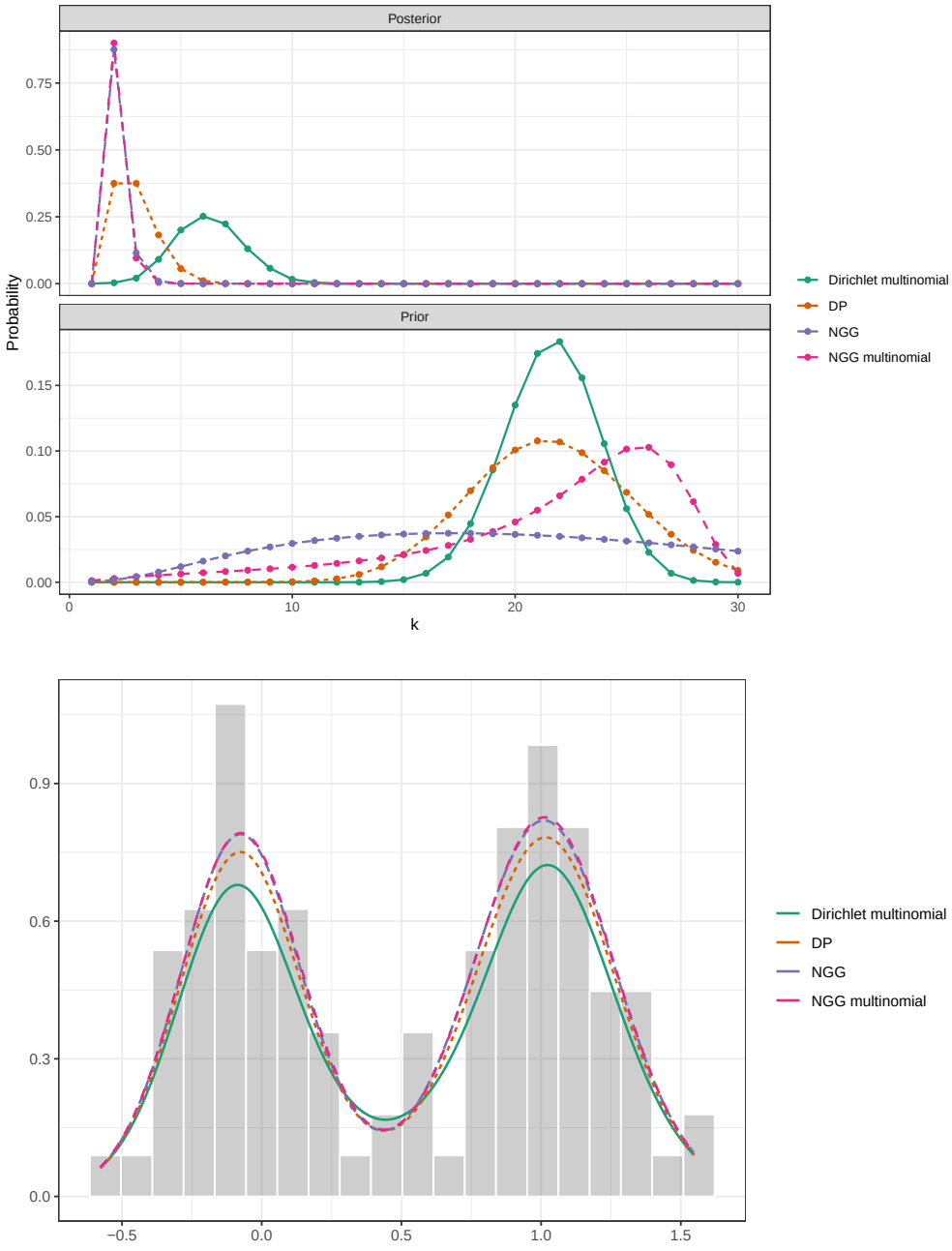

Figure 9: DATASET 2, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

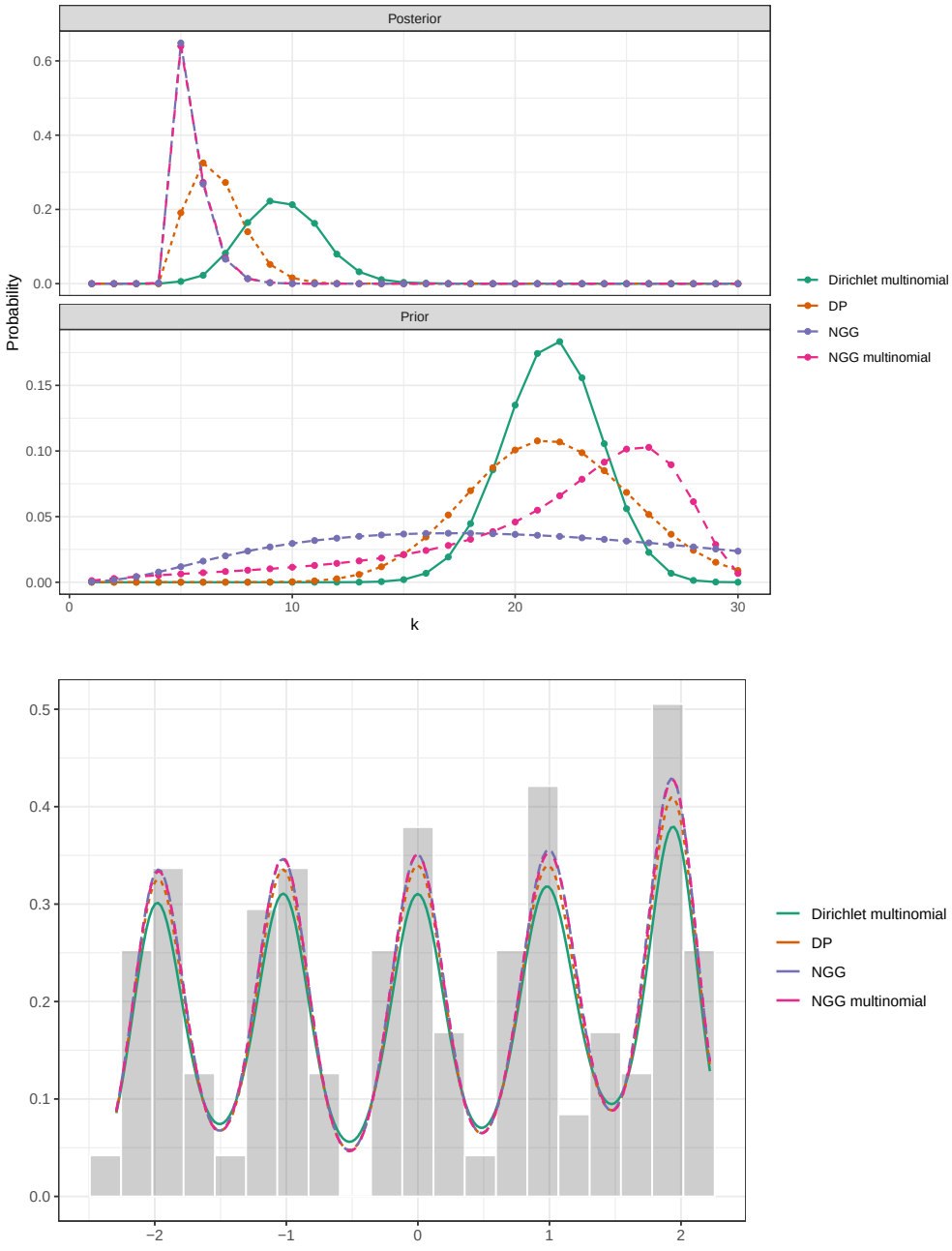

Figure 10: DATASET 3, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

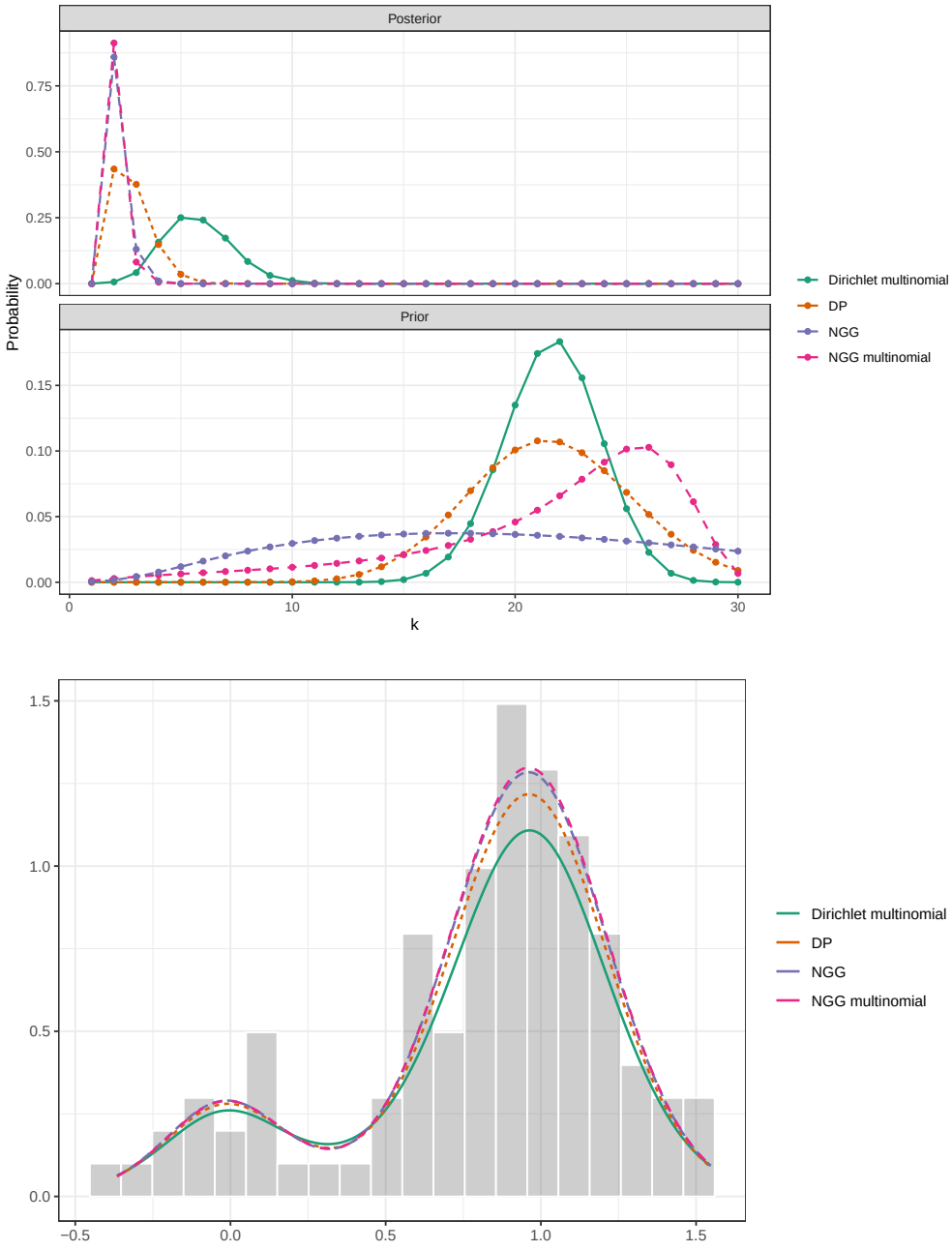

Figure 11: DATASET 4, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

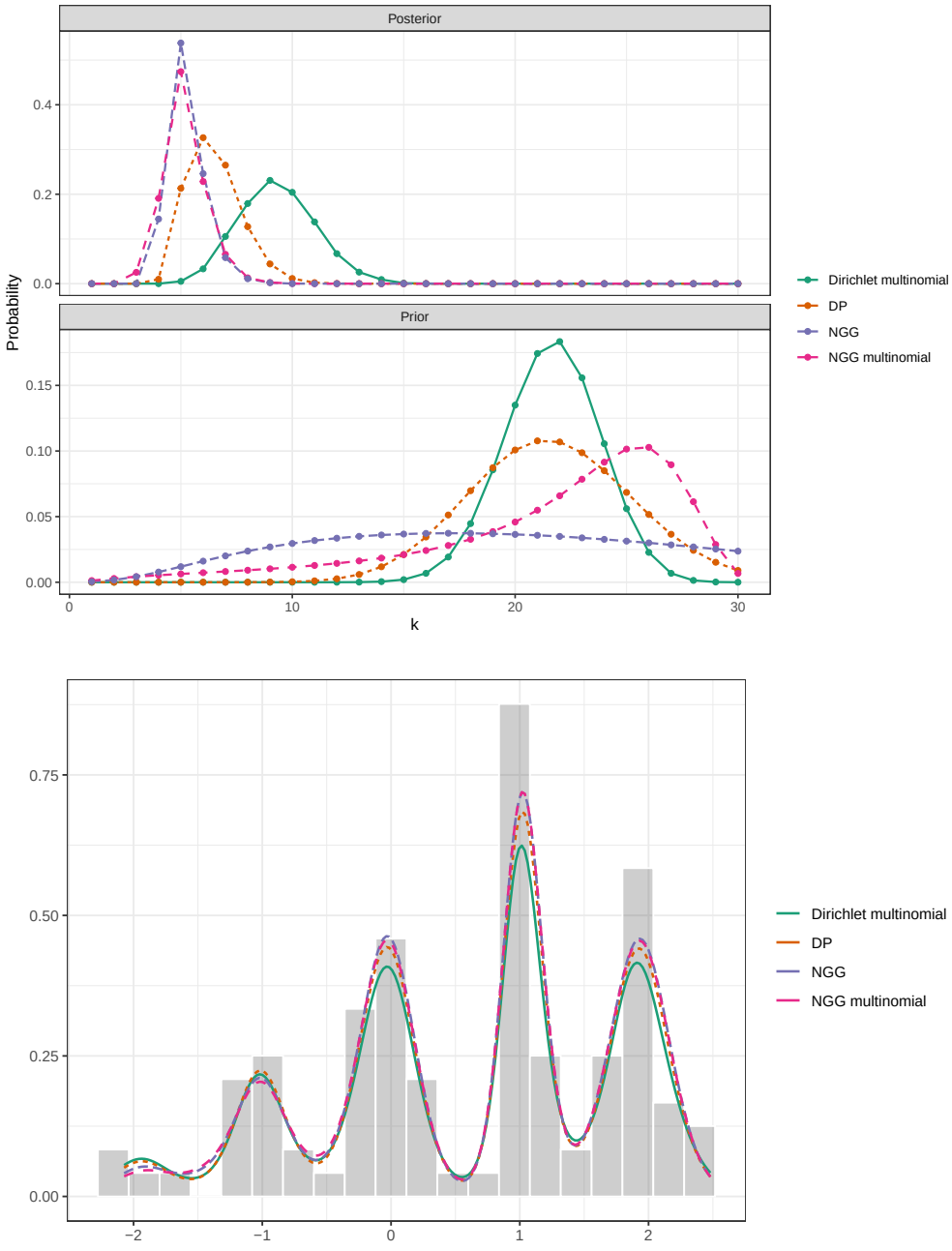

Figure 12: DATASET 5, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

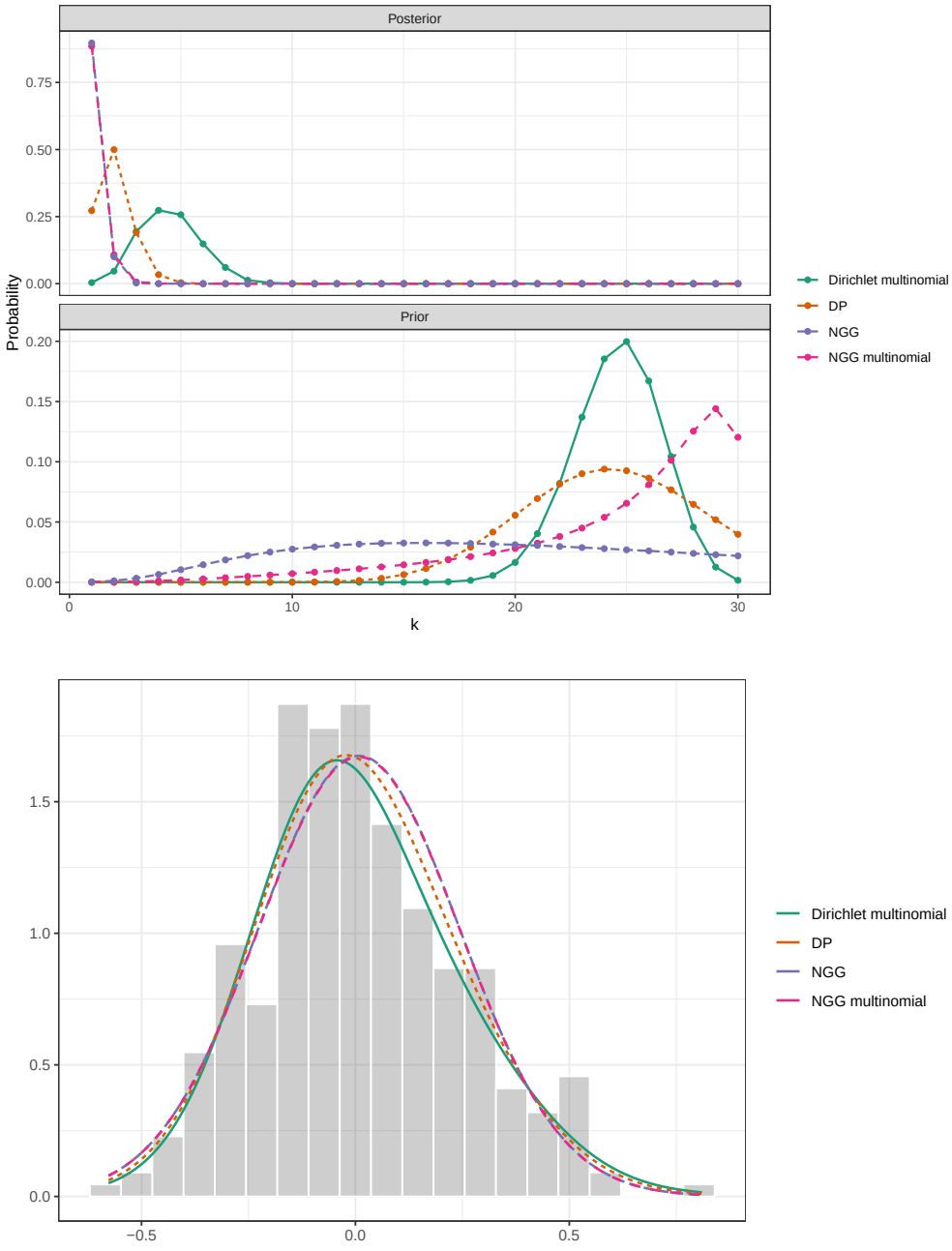

Figure 13: DATASET 1, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

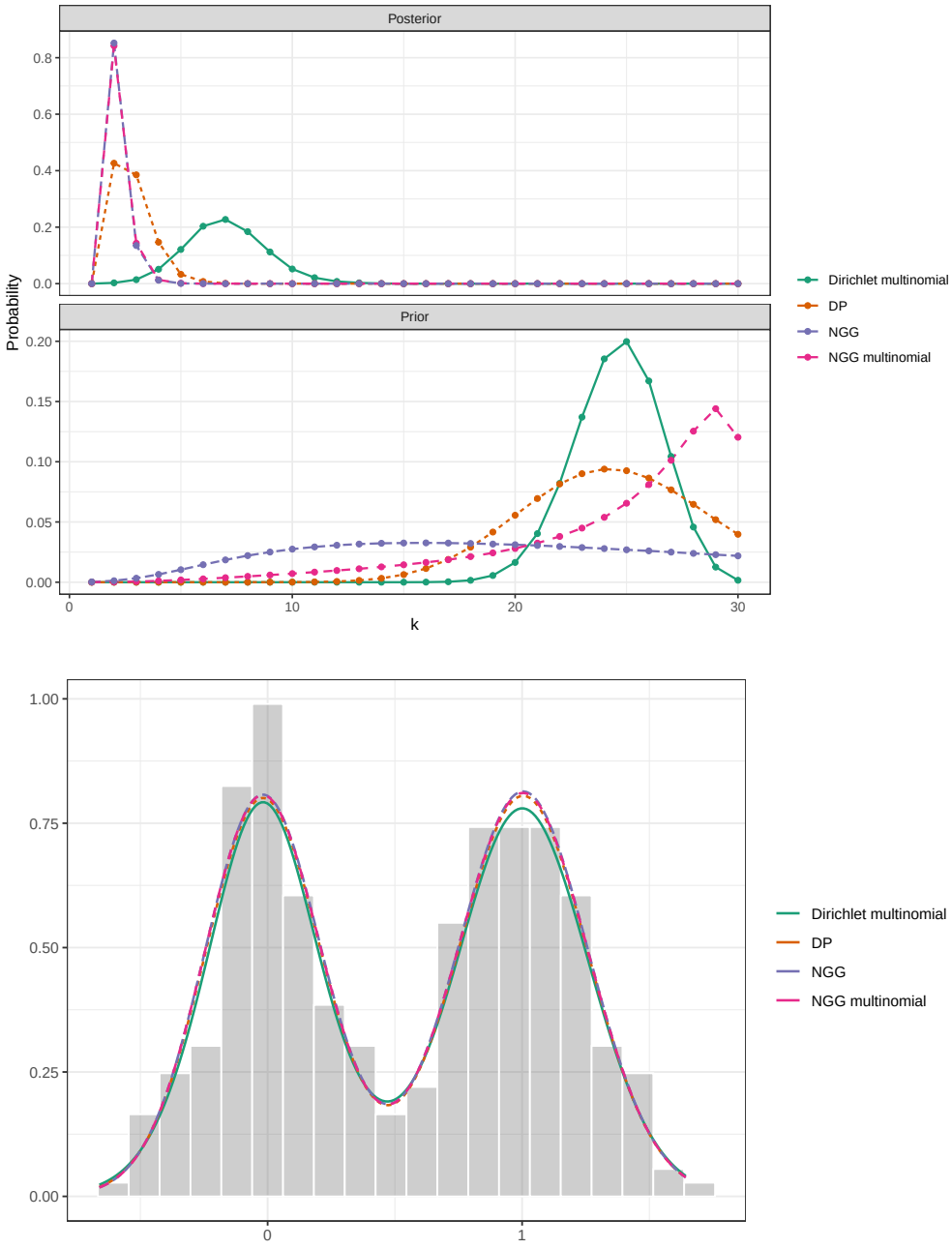

Figure 14: DATASET 2, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

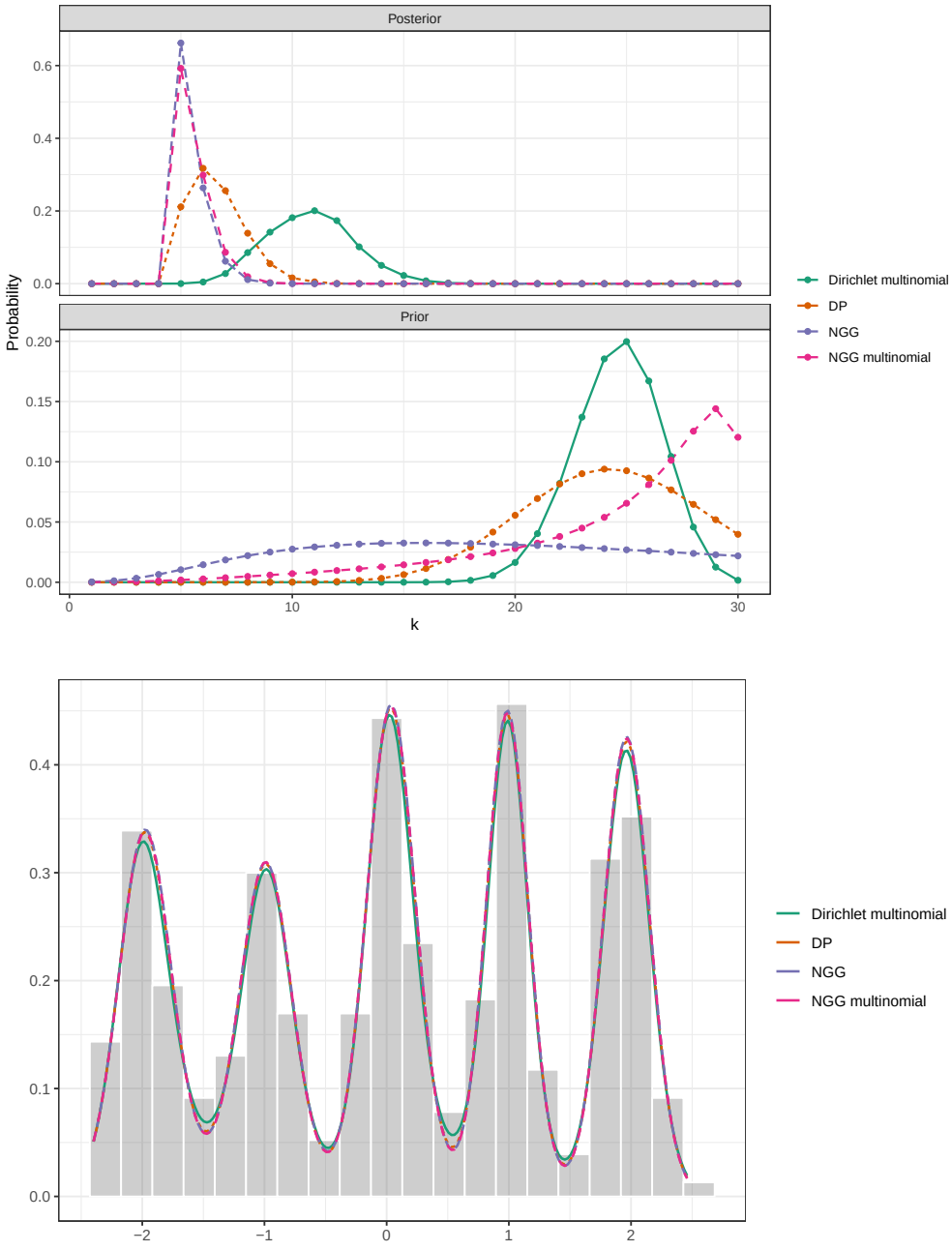

Figure 15: DATASET 3, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

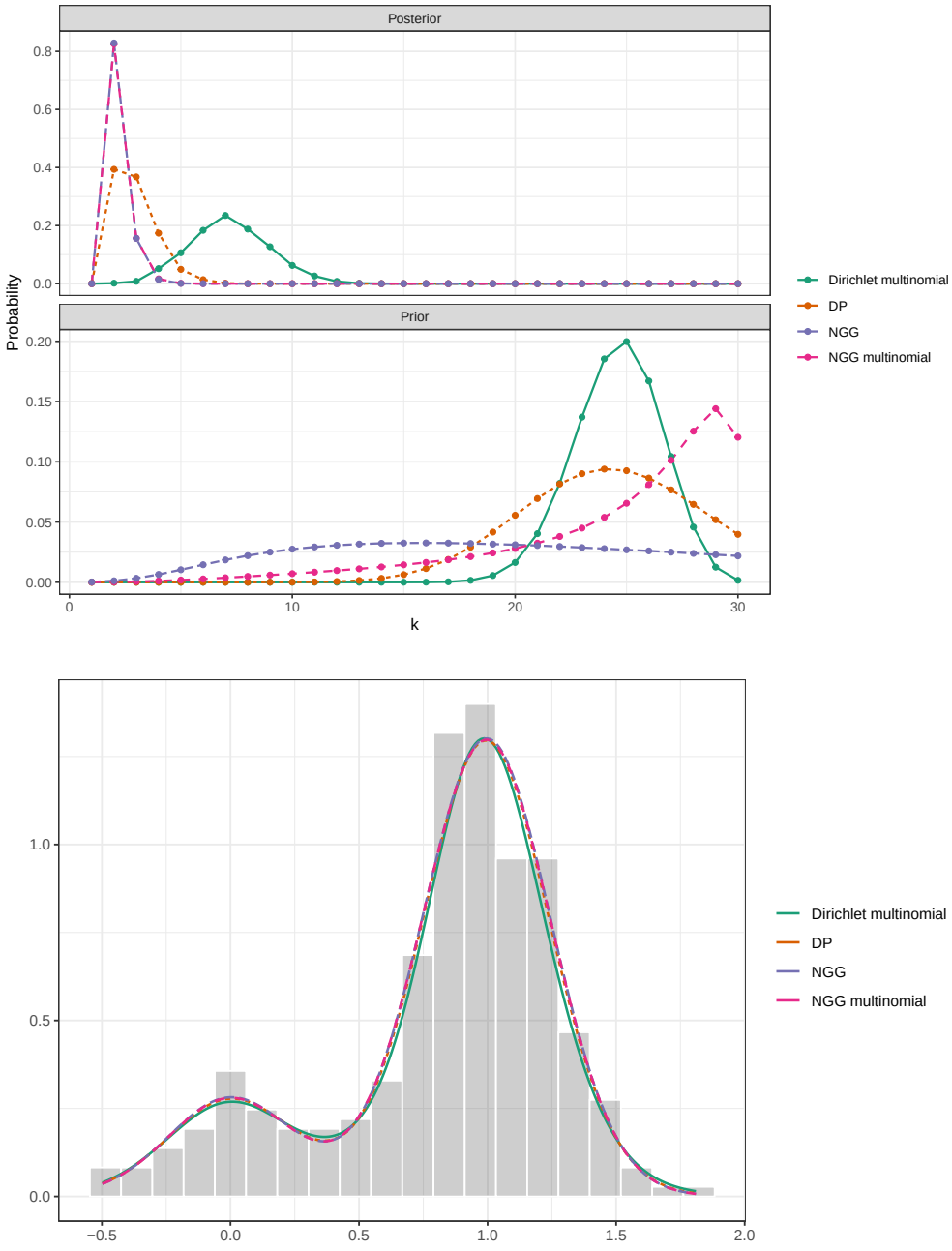

Figure 16: DATASET 4, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

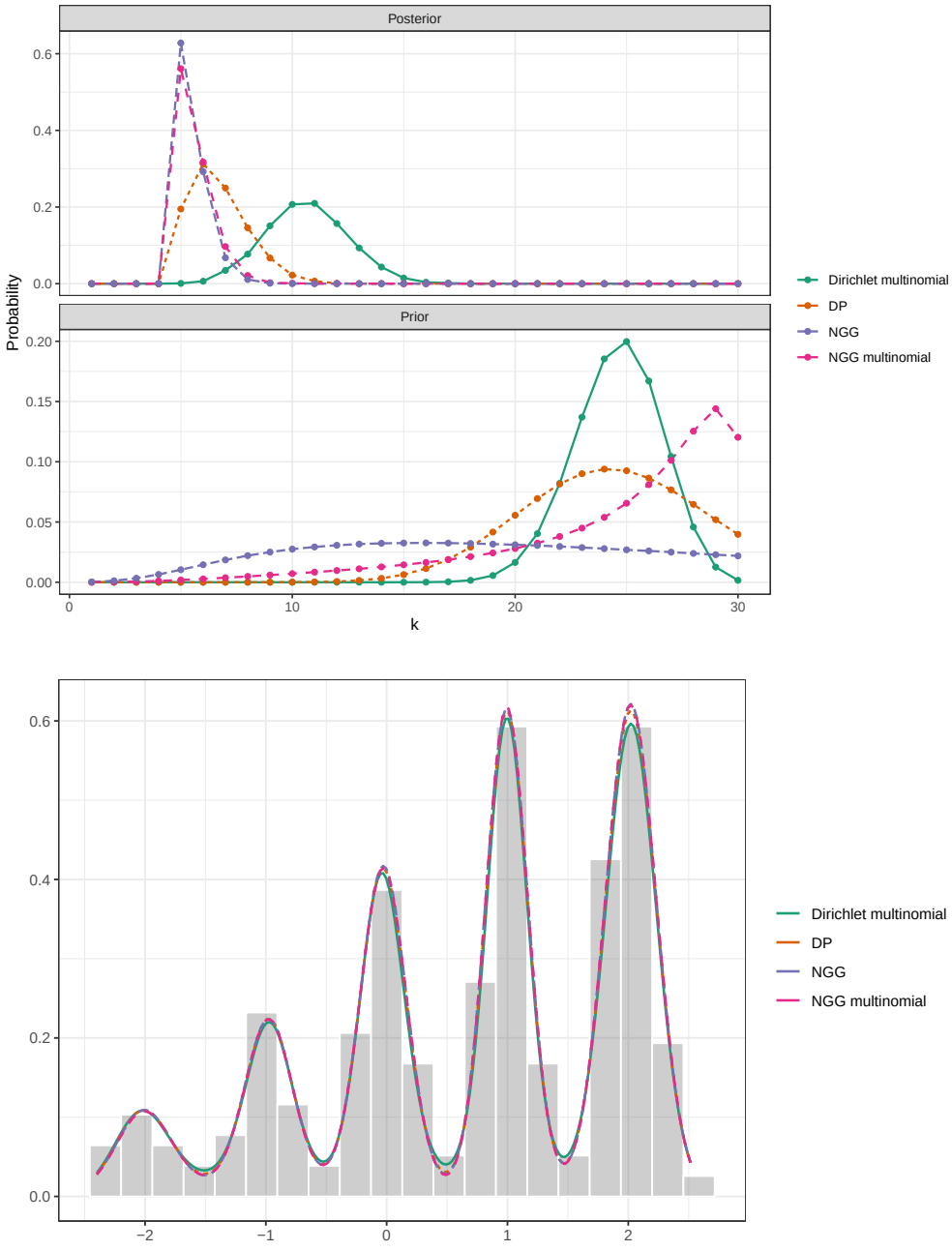

Figure 17: DATASET 5, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

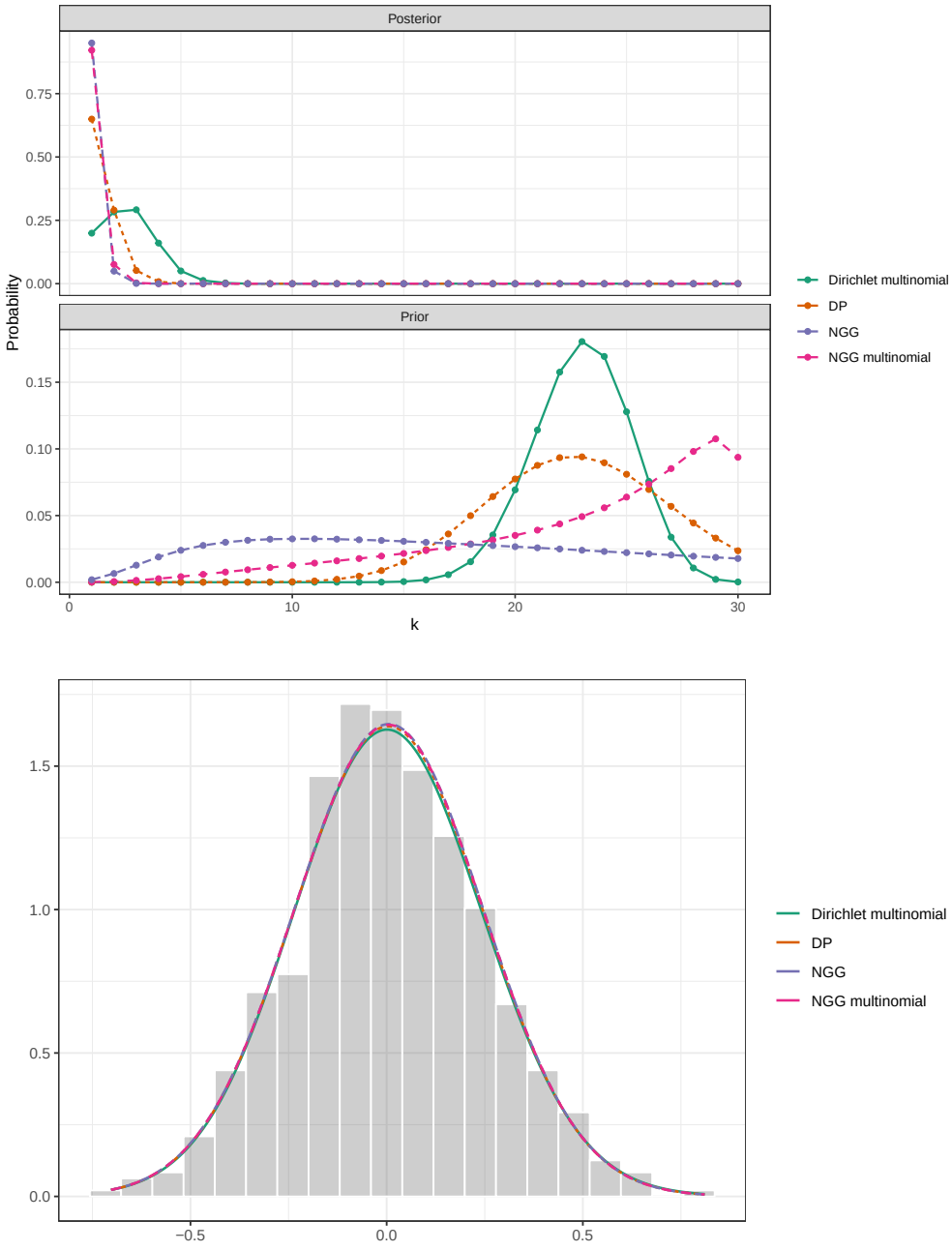

Figure 18: DATASET 1, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

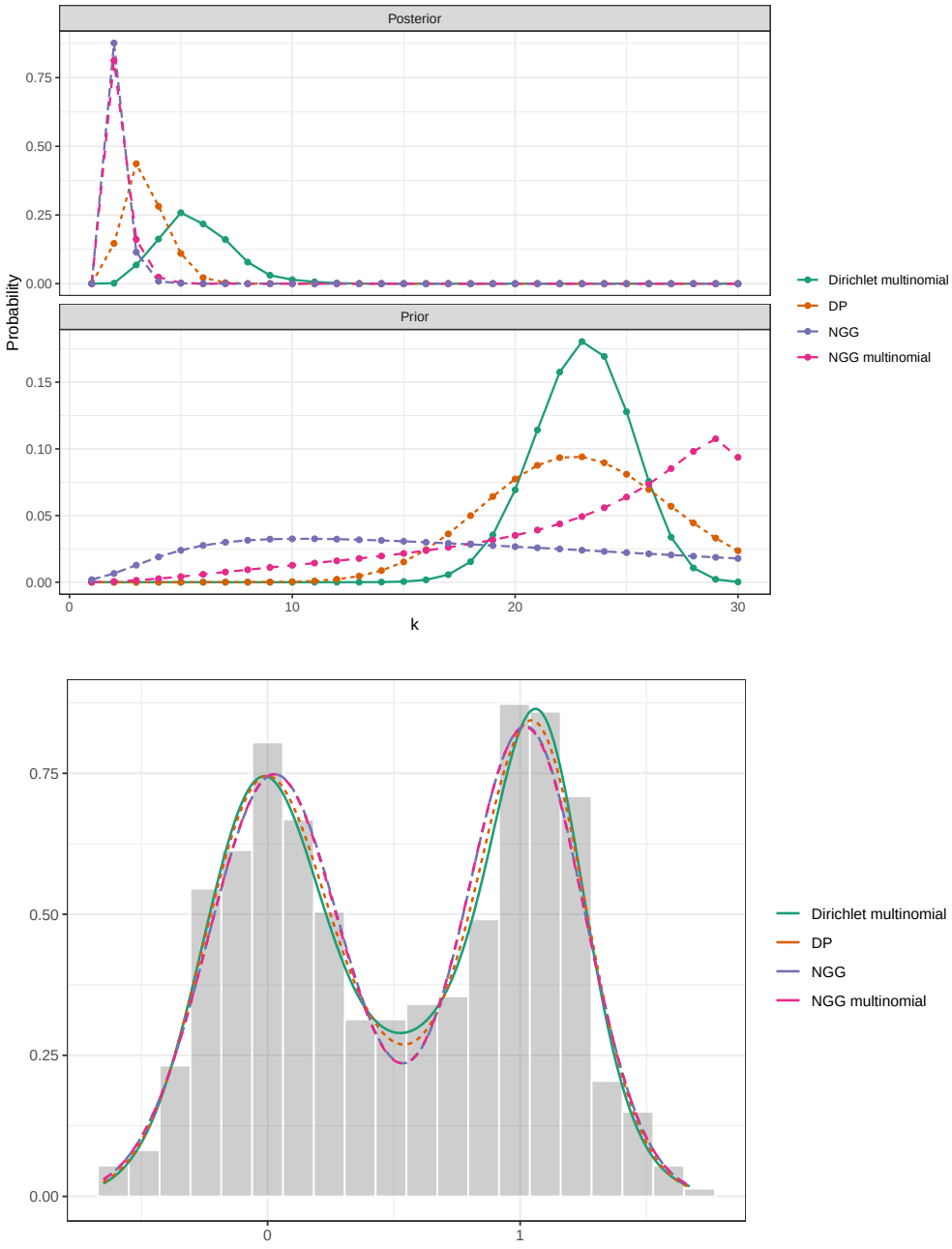

Figure 19: DATASET 2, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

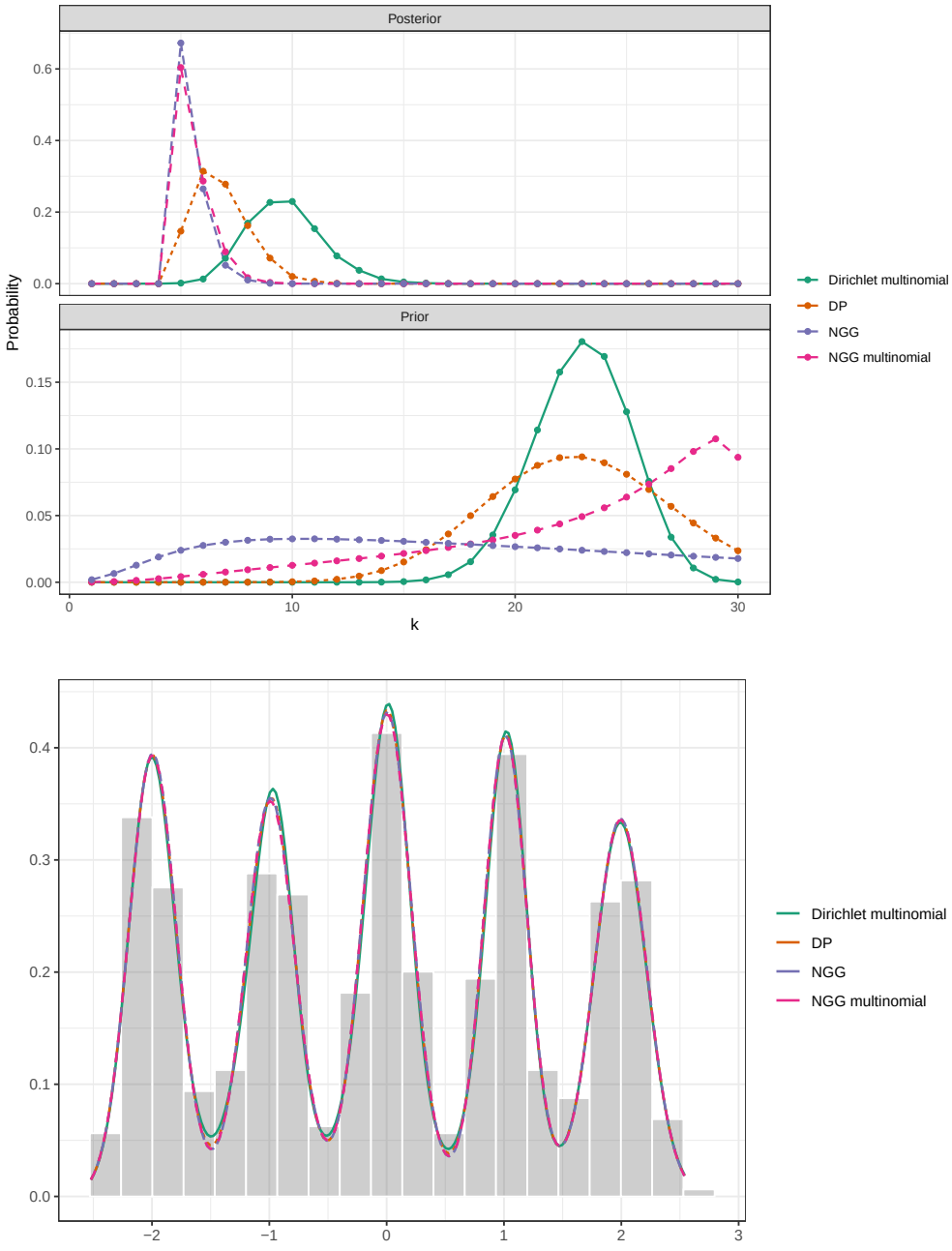

Figure 20: DATASET 3, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

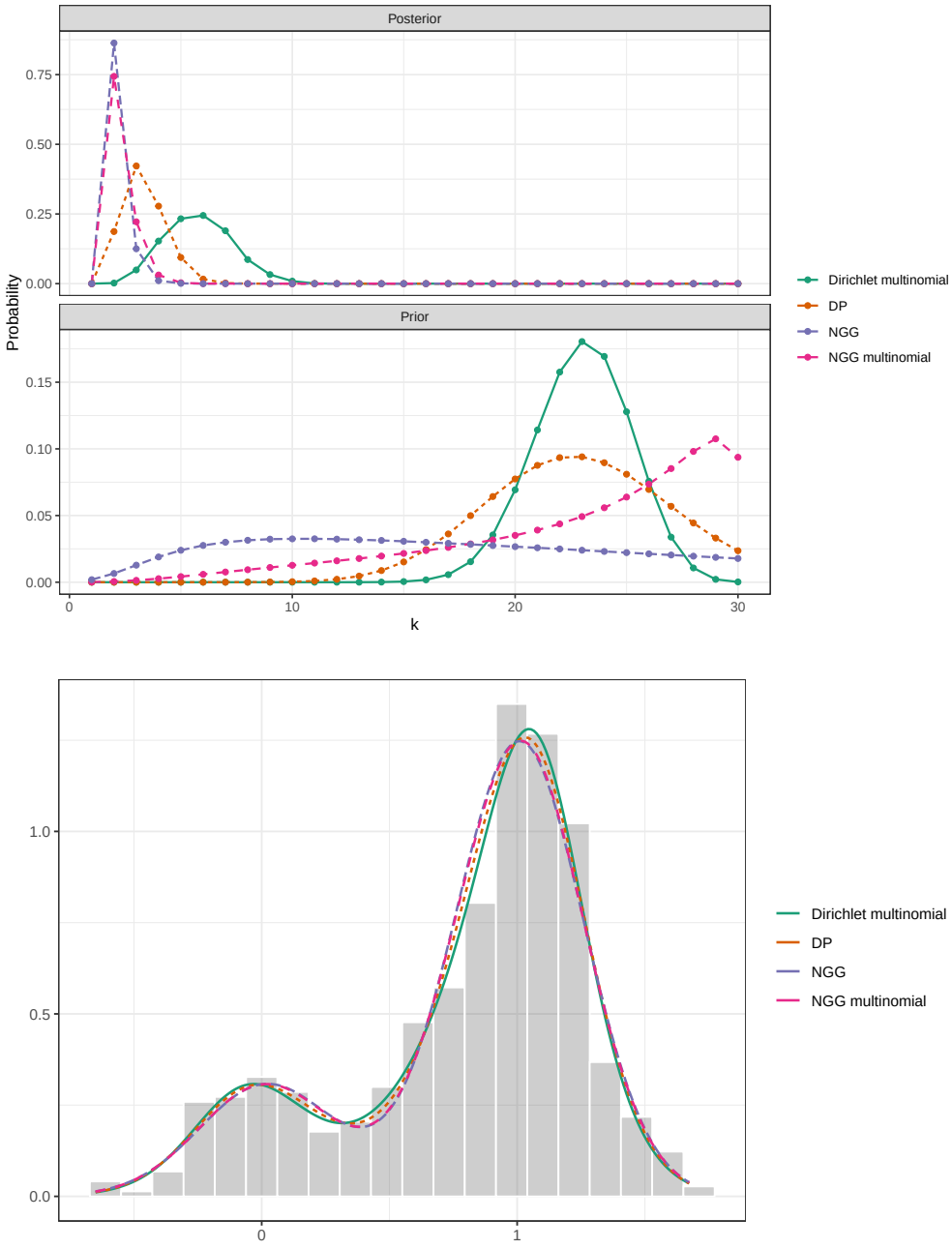

Figure 21: DATASET 4, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

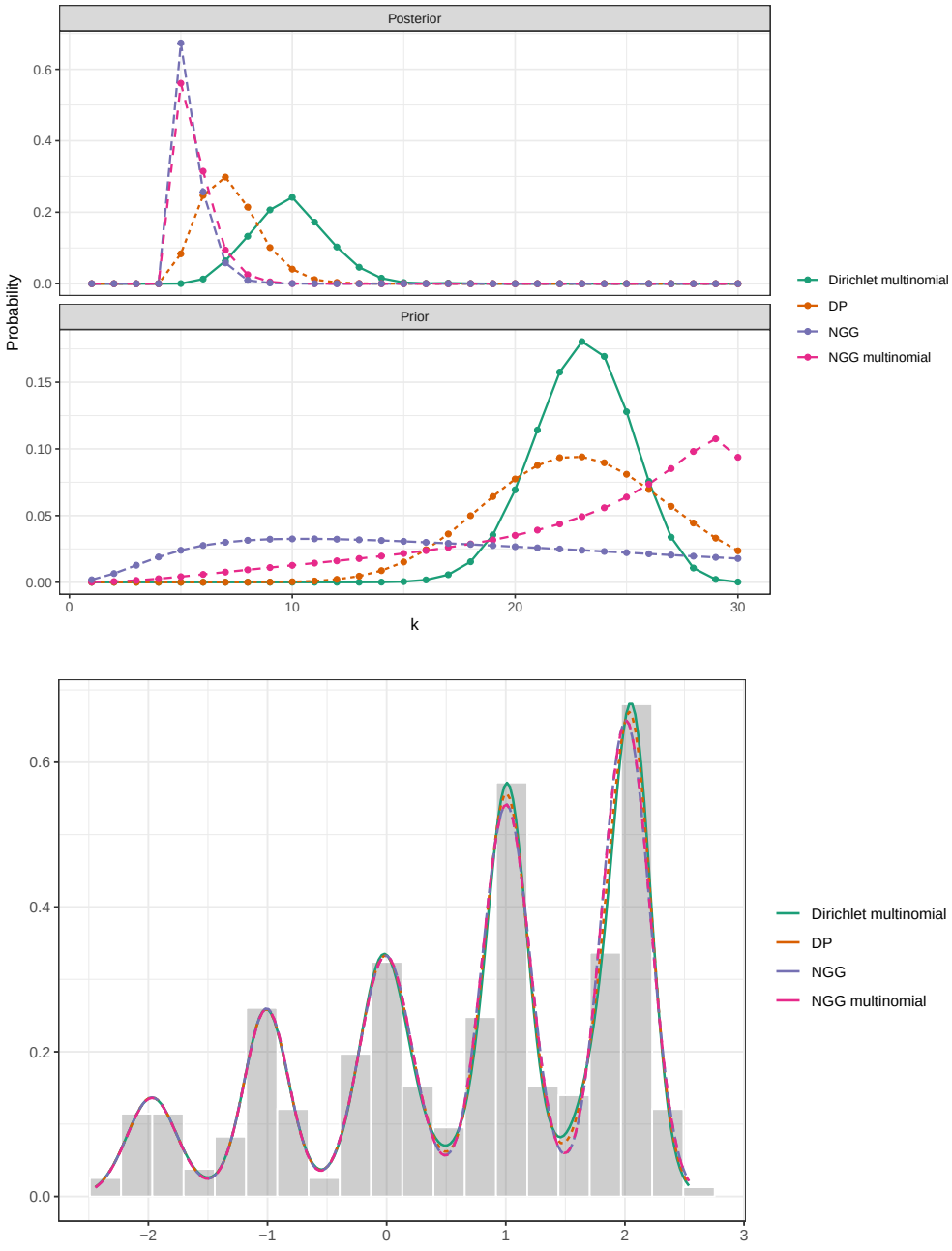

Figure 22: DATASET 5, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

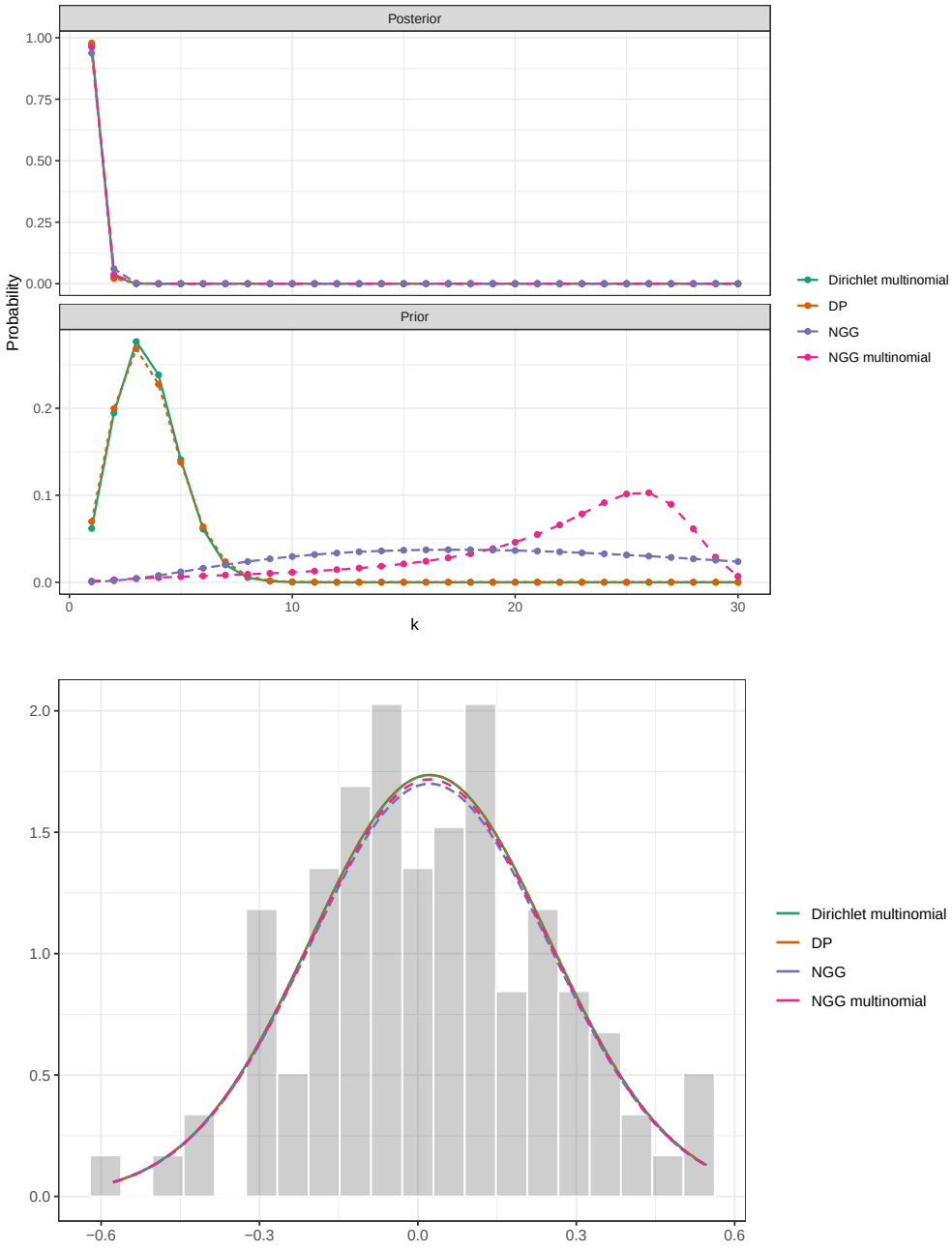

Figure 23: DATASET 1, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

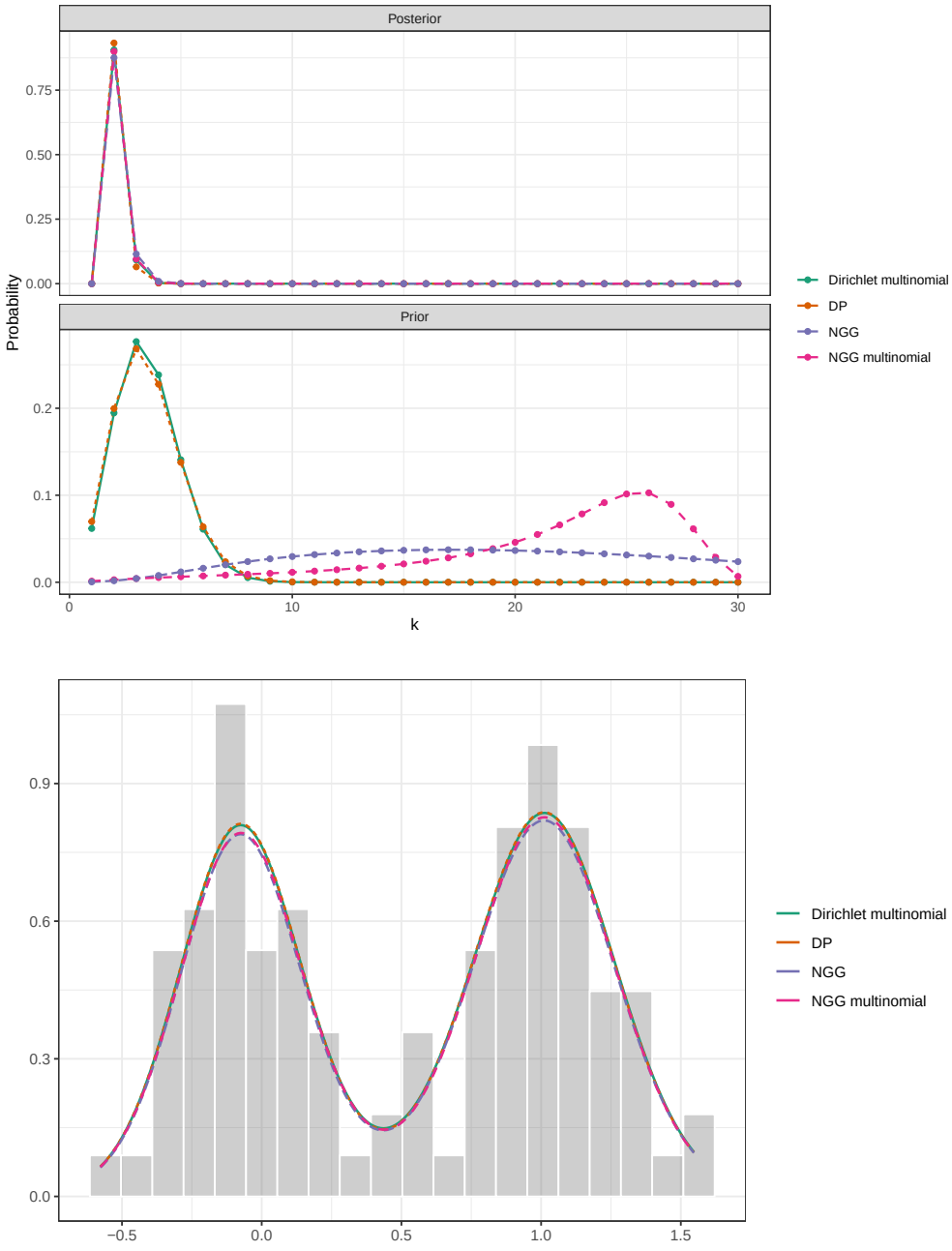

Figure 24: DATASET 2, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$ Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

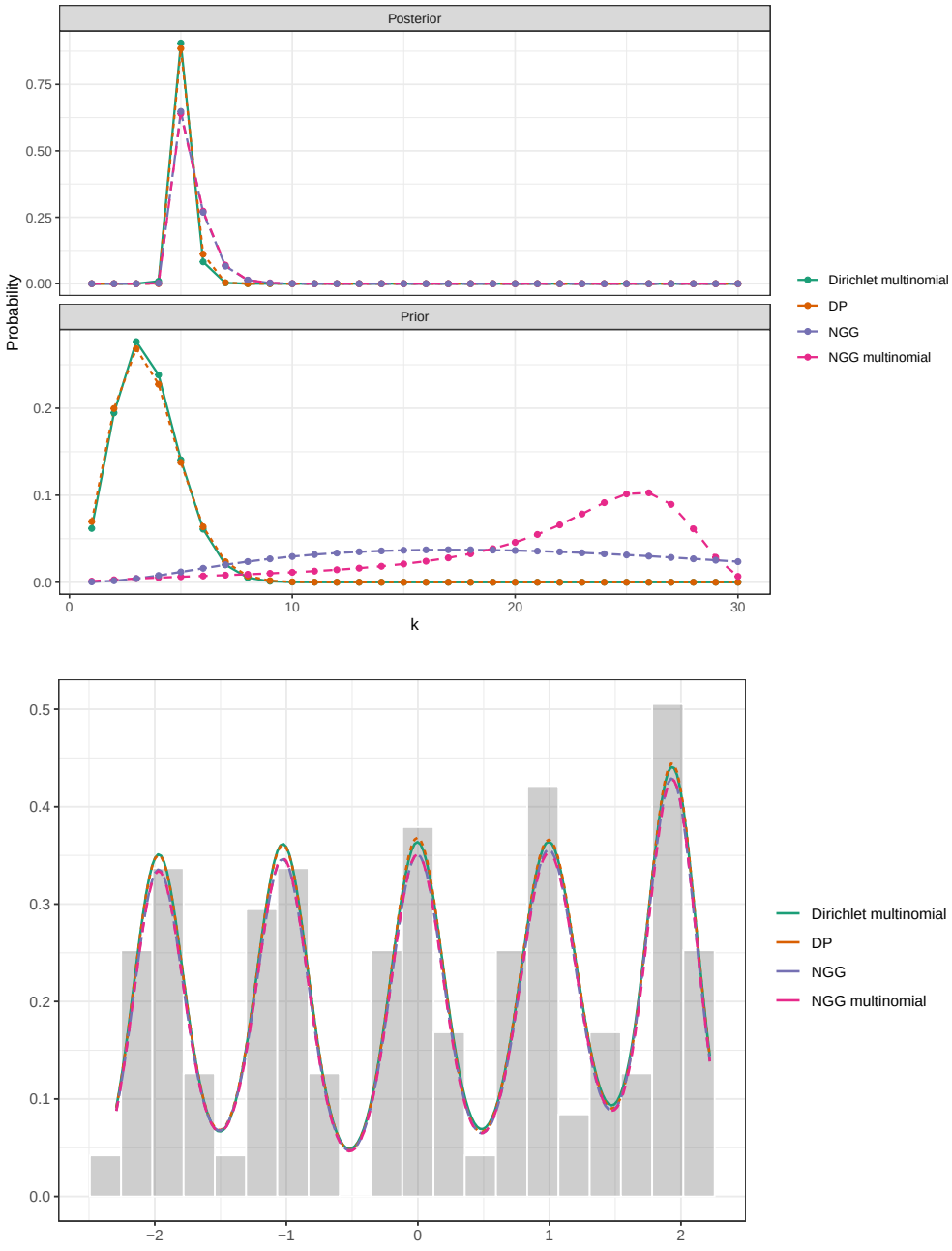

Figure 25: DATASET 3, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

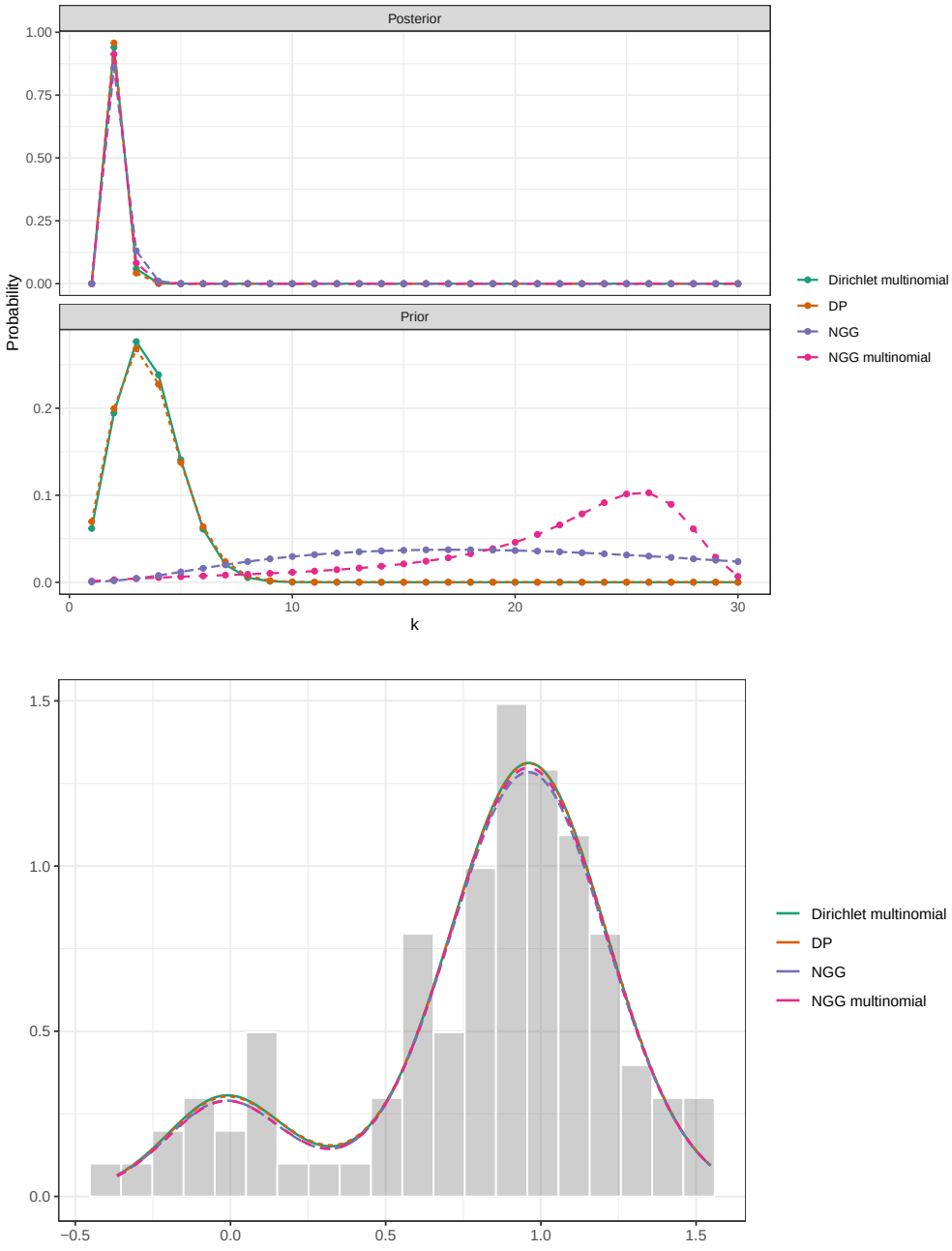

Figure 26: DATASET 4, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

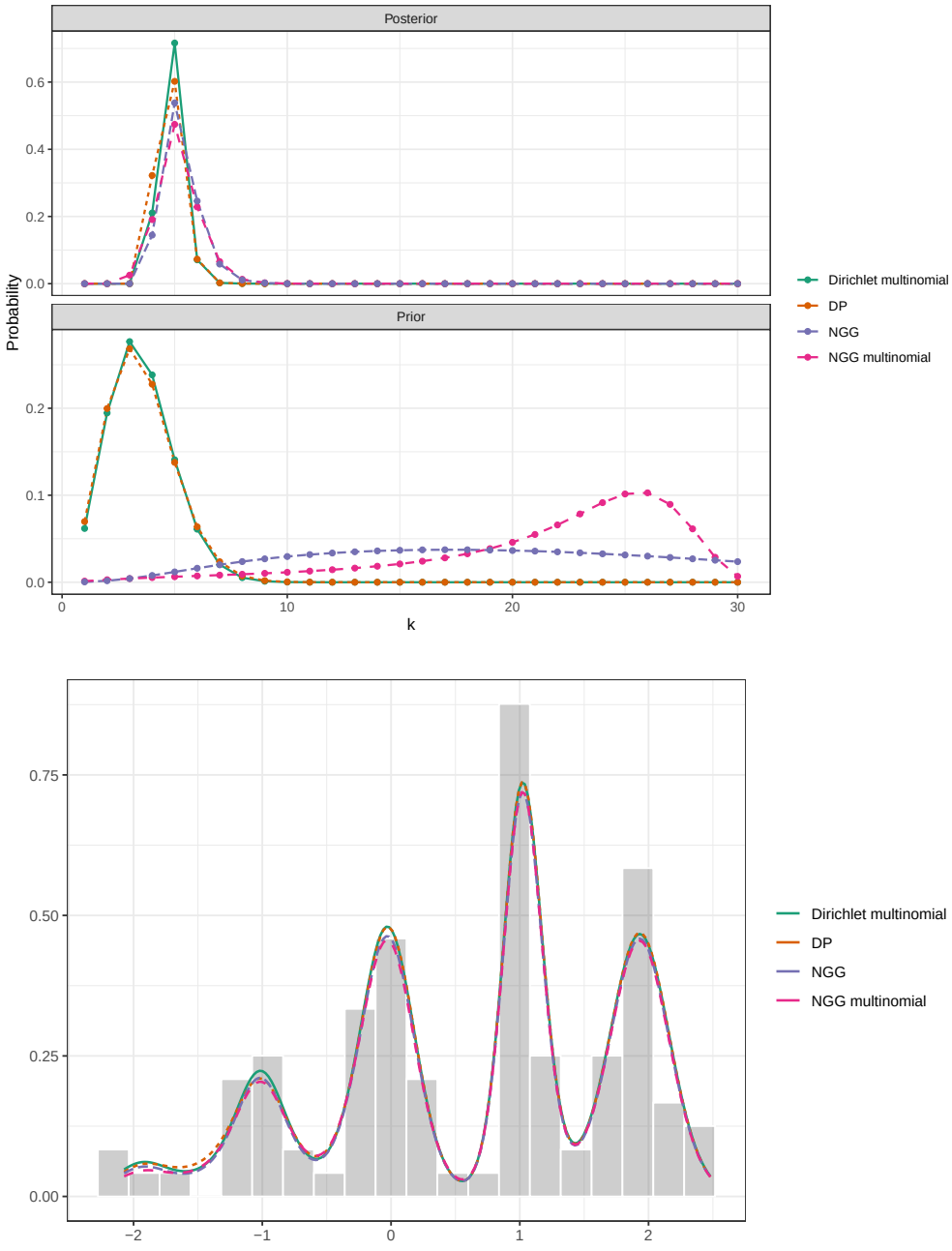

Figure 27: DATASET 5, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.