SUPPLEMENTARY MATERIAL of 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 and Statistics, University of Milano-Bicocca

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 A.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*

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

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

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

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

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

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

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

from which the result follows. As for the covariance, note that $\text{Var}(\tilde{p}_H(A_1), \tilde{p}_H(A_2)) =$

$P(A)(1-P(A))\mathbb{E}\left(\sum_{h=1}^{H}\pi_h^2\right)$ 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 $m \geq 1$

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

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 (A.1) 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 (A.1), 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

$$\mathbb{P}(K_{n,H} = k) = \sum_{t=k}^{n} \mathbb{P}(K_{n,\infty} = t)\mathbb{P}(K_{t,0} = k), \tag{A.2}$$

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 (A.2). 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 A.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 A.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 A.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 A.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 (A.1), 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 A.2 with the above Laplace functional the result follows.

A.9 Proof of Corollary 4

By exploiting equation (A.1), 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 A.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

$$\begin{aligned}
&\frac{1}{\ell_j!}\sum_{\boldsymbol{q}_j}\binom{n_j}{q_{j1},\ldots,q_{j\ell_j}}\Pi_\infty(q_{11},\ldots,q_{1\ell_1},\ldots,q_{k1},\ldots,q_{k\ell_k}) = \\
&\quad = \mathscr{V}_{n,|\boldsymbol{\ell}|}\frac{1}{\ell_j!}\sum_{\boldsymbol{q}_j}\binom{n_j}{q_{j1},\ldots,q_{j\ell_j}}\prod_{r=1}^{\ell_j}(1-\sigma)_{q_{jr}-1} = \mathscr{V}_{n,|\boldsymbol{\ell}|}\frac{\mathscr{C}(n_j,\ell_j;\sigma)}{\sigma^{\ell_j}},
\end{aligned}$$

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 (A.1).

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

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

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 1, 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 1 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 2-4 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.

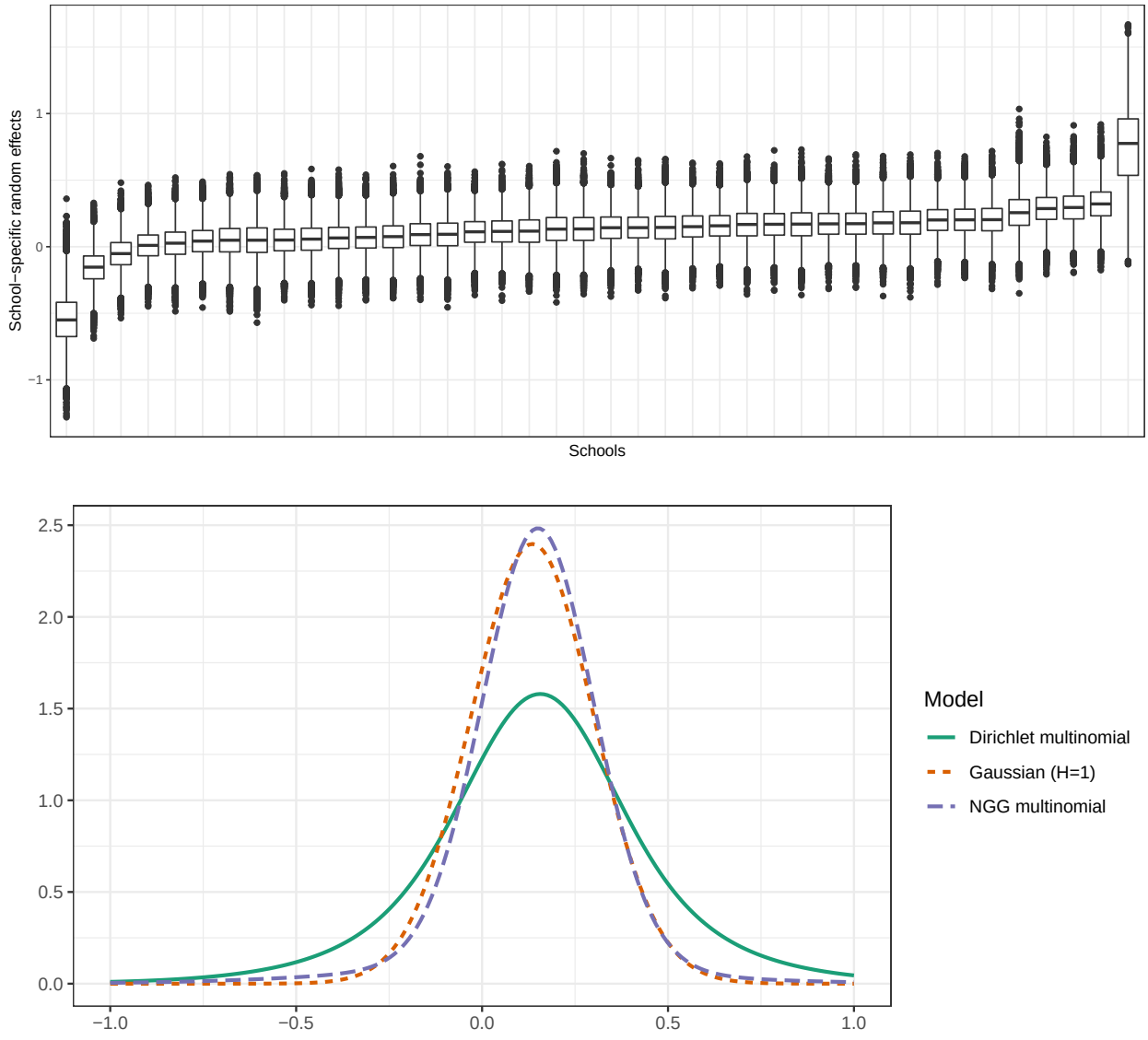

Figure 1: Top panel: Posterior distribution of the random effects η_j for 40 randomly selected schools, ordered according to the posterior median. Bottom panel: Posterior mean of the random density $\sum_{h=1}^{H} \pi_h \mathcal{N}(\bar{\mu}_h, \bar{\sigma}_h^2)$.

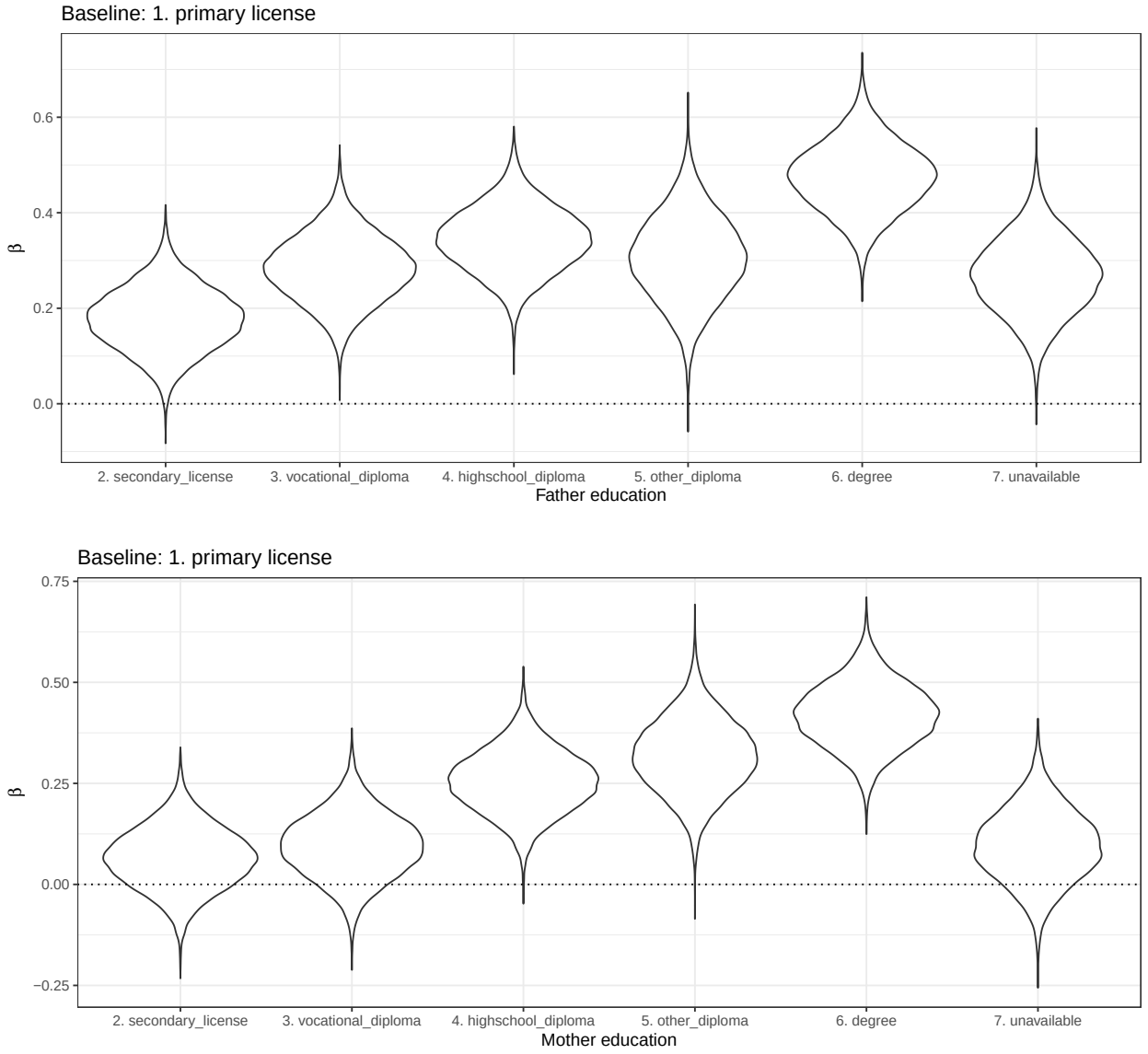

Figure 2: Posterior distribution of the $\boldsymbol{\gamma}$ coefficients in the INVALSI application.

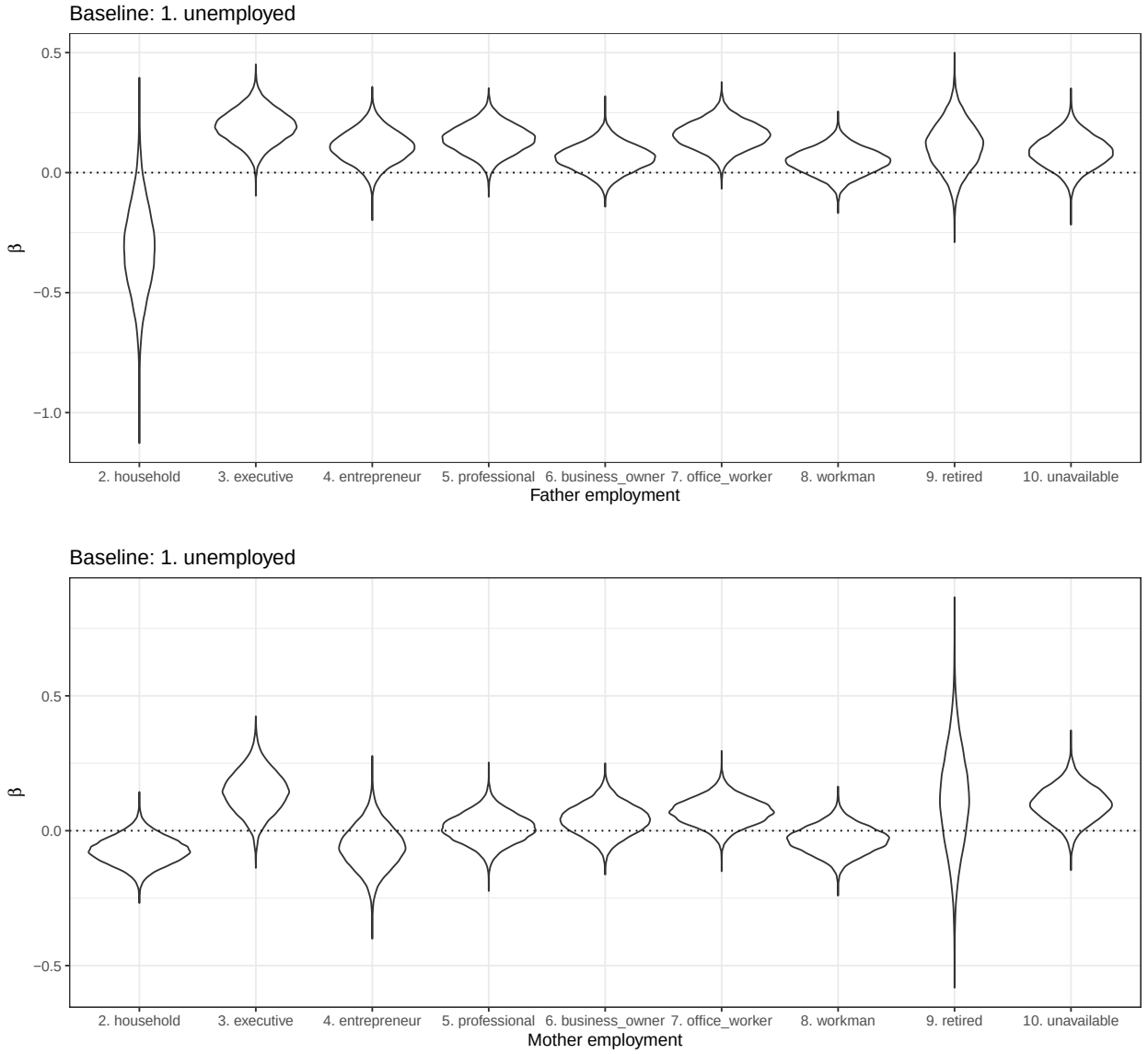

Figure 3: Posterior distribution of the $\boldsymbol{\gamma}$ coefficients in the INVALSI application.

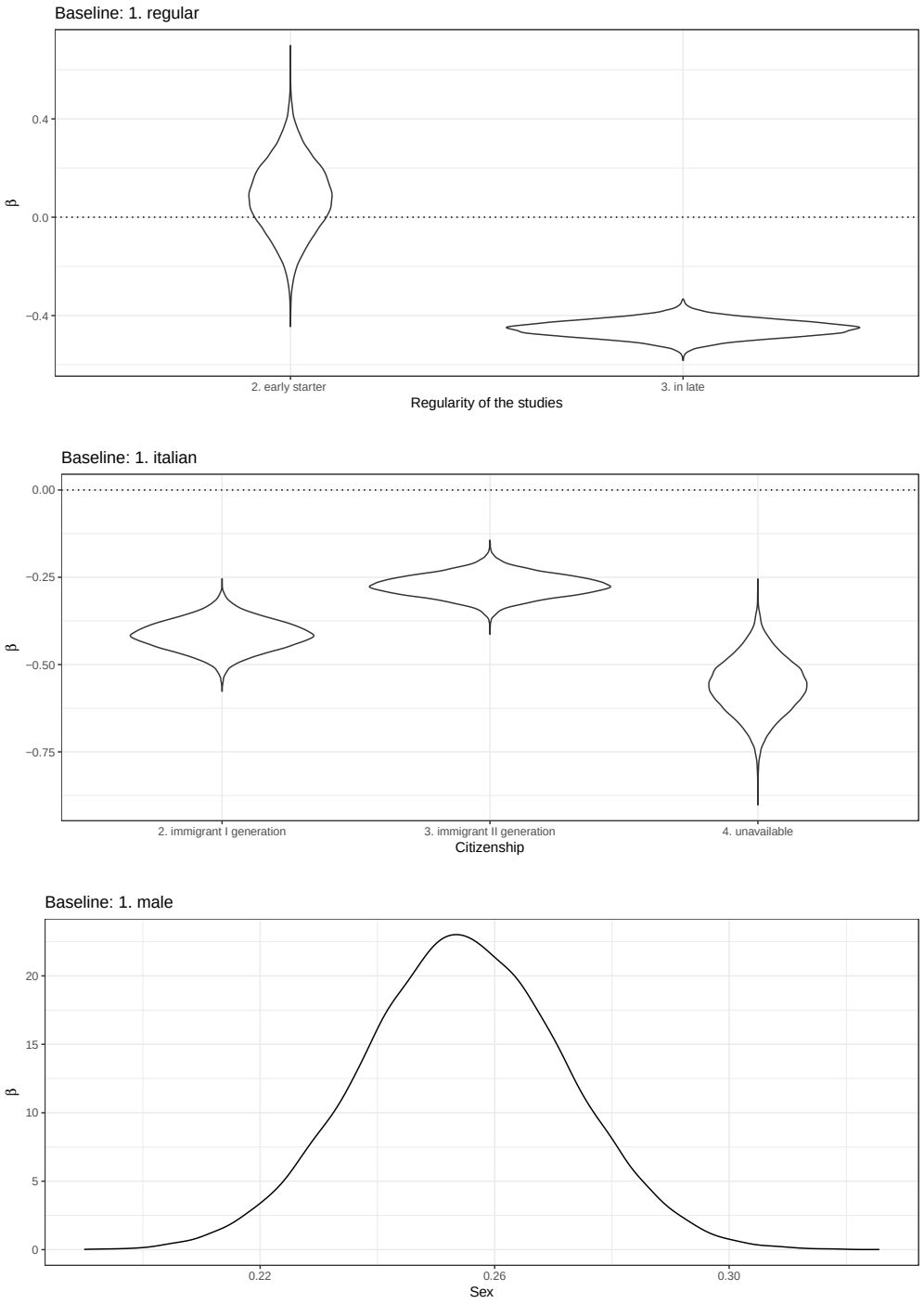

Figure 4: Posterior distribution of the $\boldsymbol{\gamma}$ coefficients in the INVALSI application.

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 1, Table 2, Table 3, and Table 4.

D.2 Summary of the results

The results of the simulation studies are reported in Figures 5–24, displaying the a priori and the a posteriori distribution of the number of clusters, together with the posterior mean of the density.

Table 1: 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 2: 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

Table 3: 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 4: 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

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 7) 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 5 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.

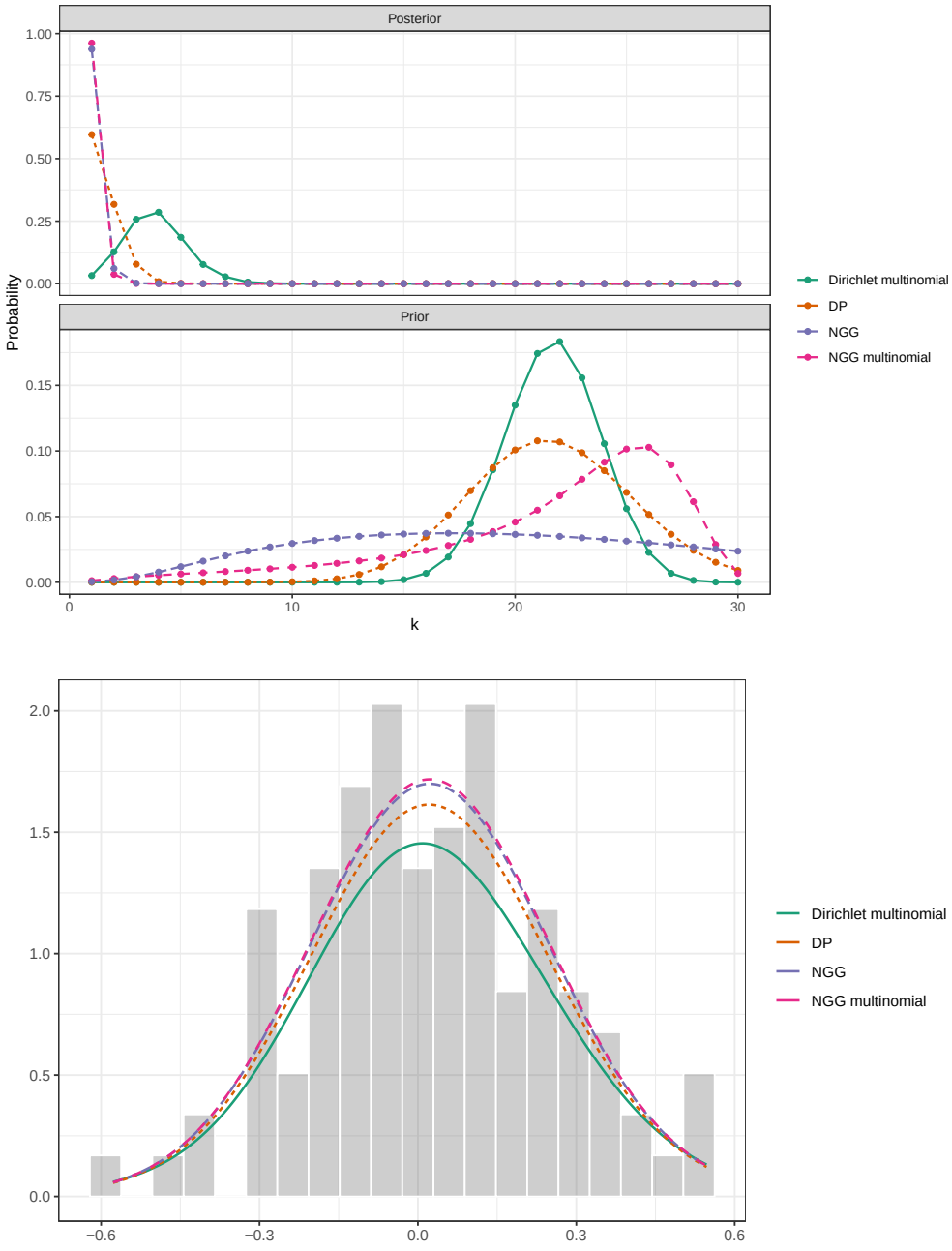

Figure 5: DATASET 1, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

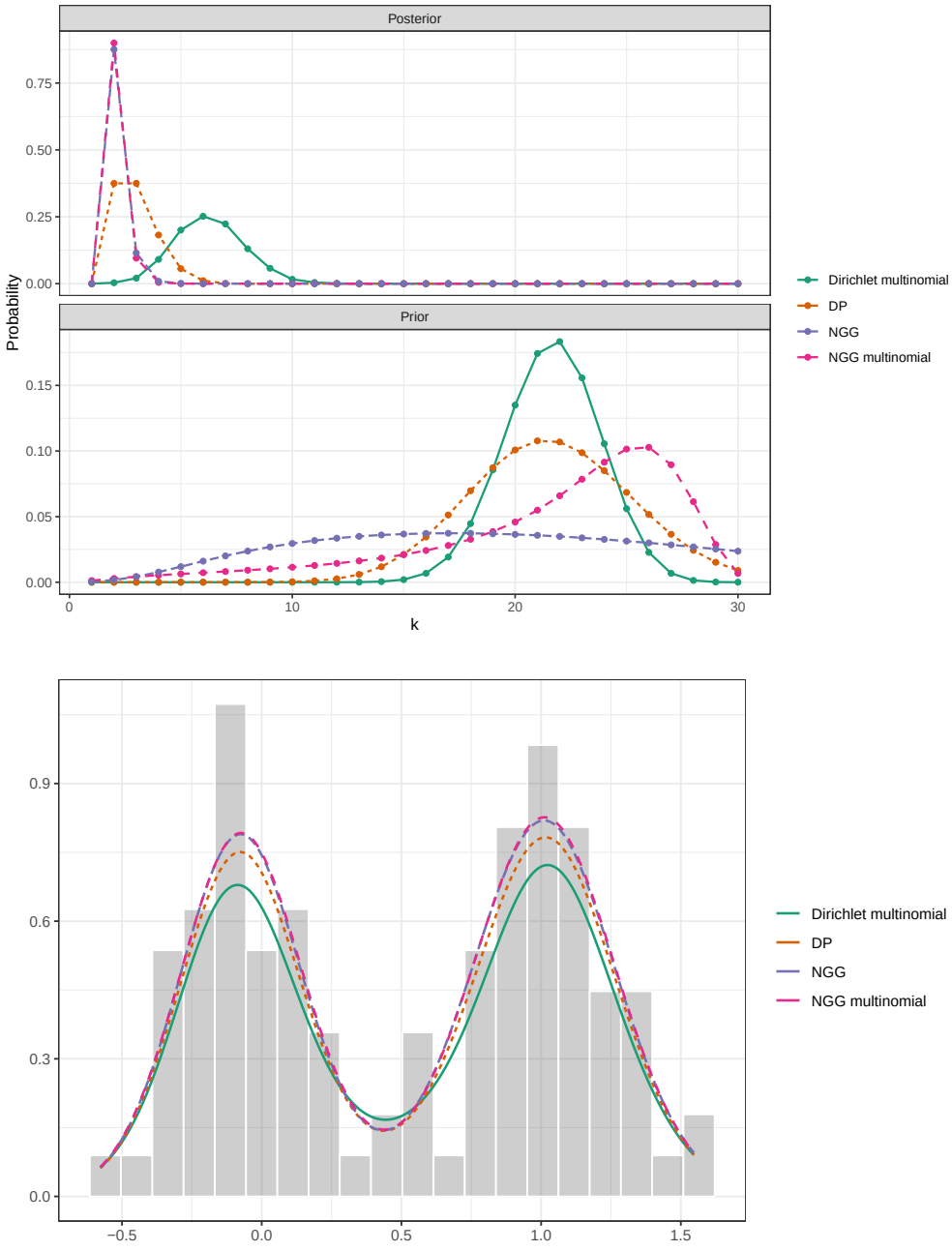

Figure 6: DATASET 2, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

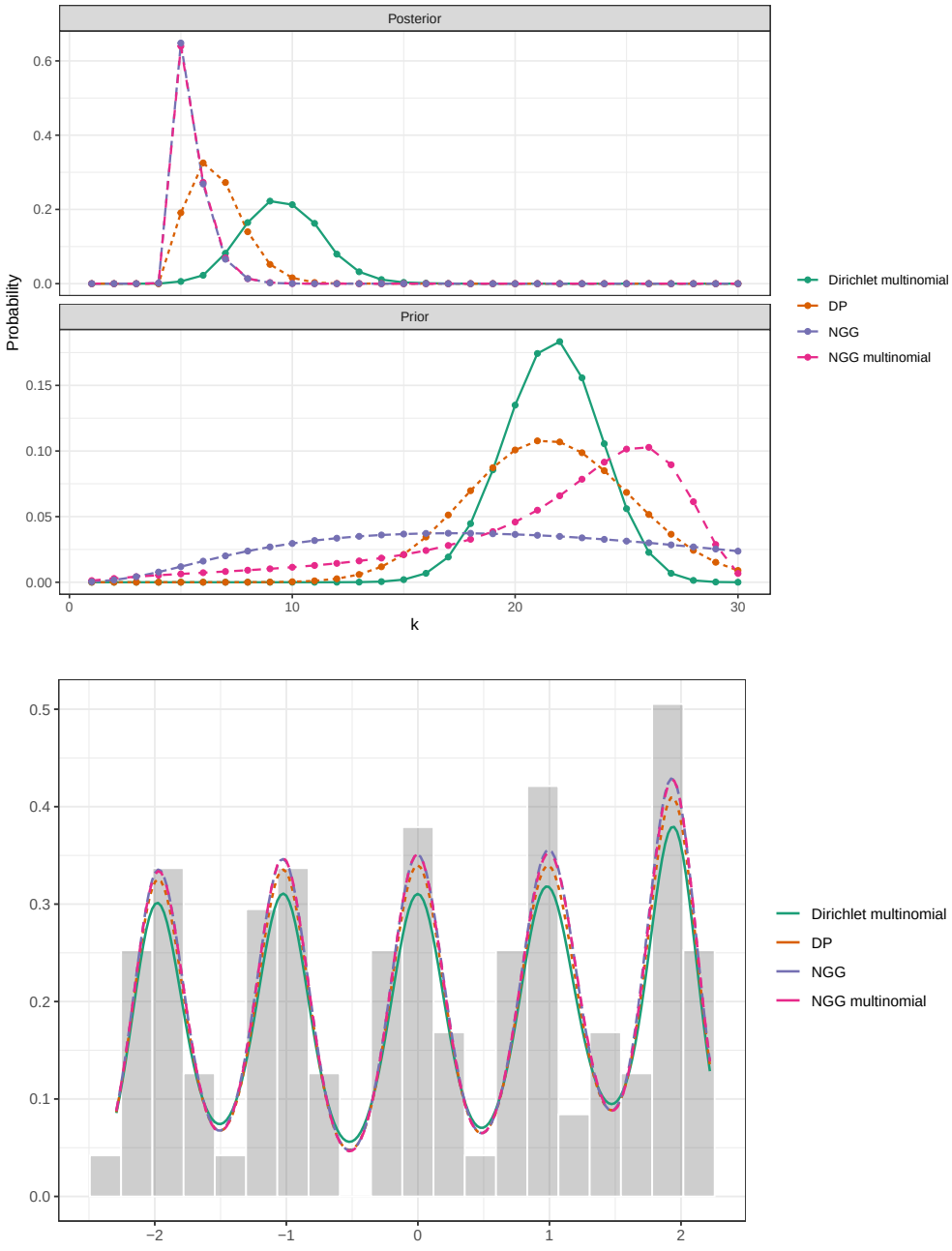

Figure 7: DATASET 3, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

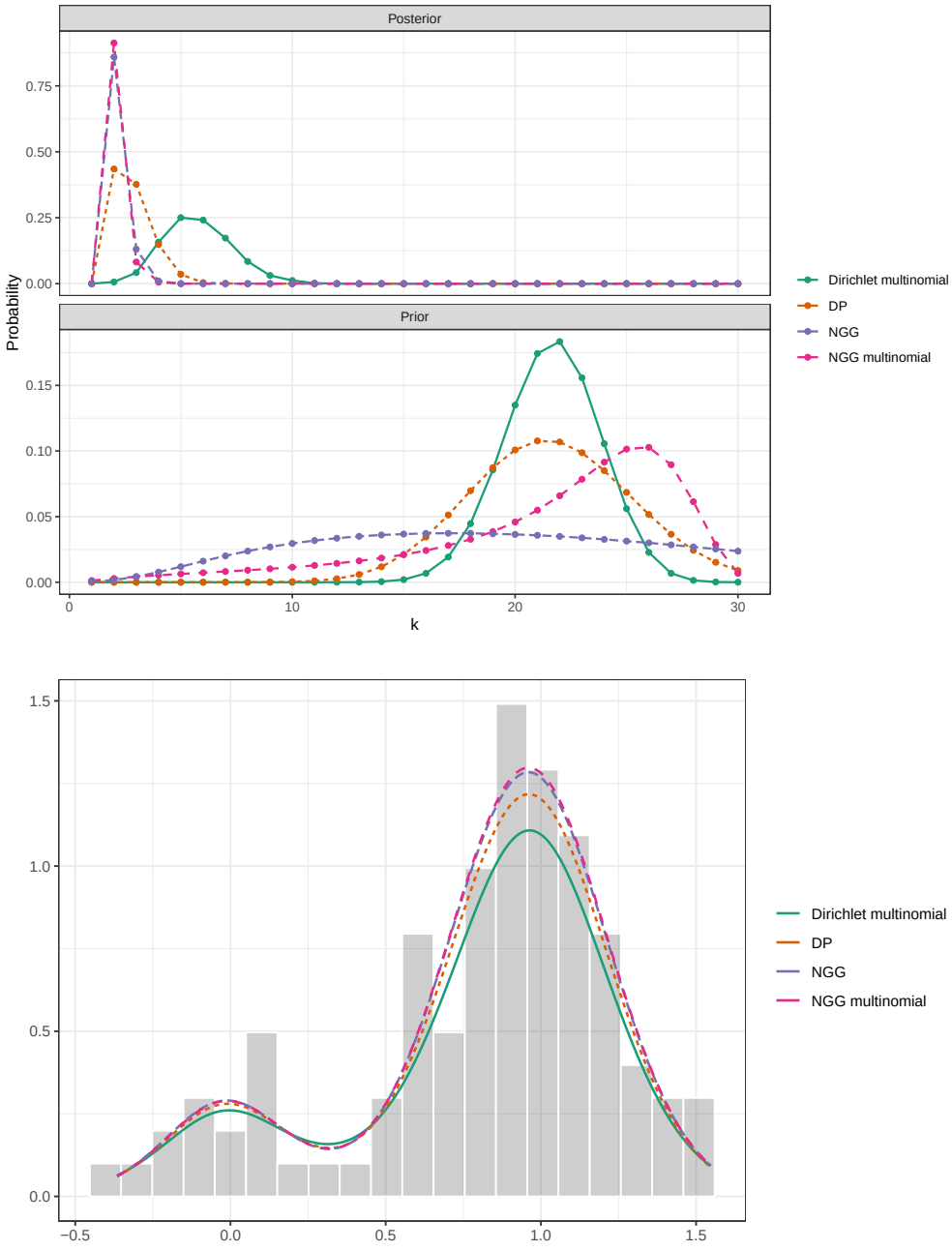

Figure 8: DATASET 4, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

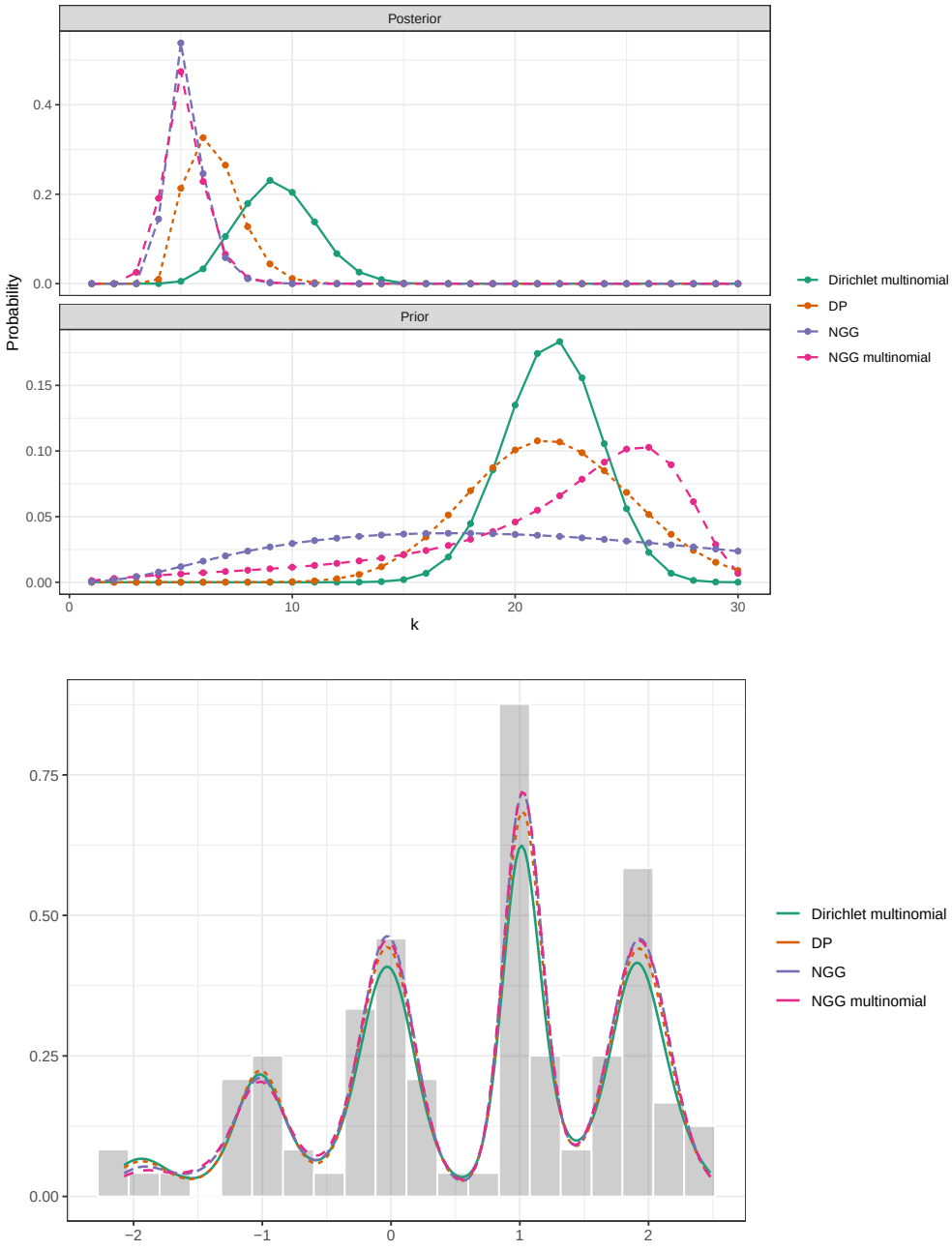

Figure 9: DATASET 5, SCENARIO A. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

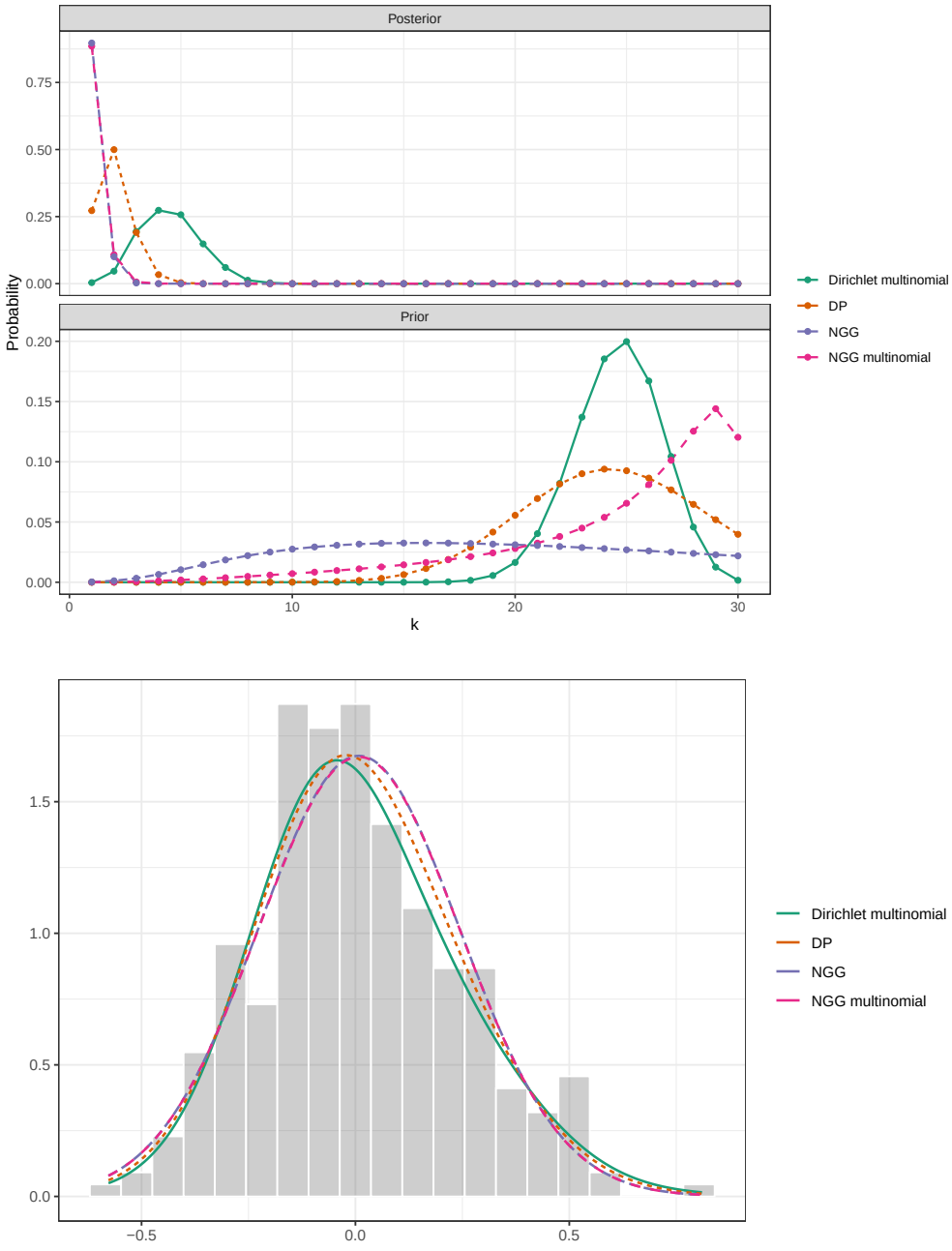

Figure 10: DATASET 1, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

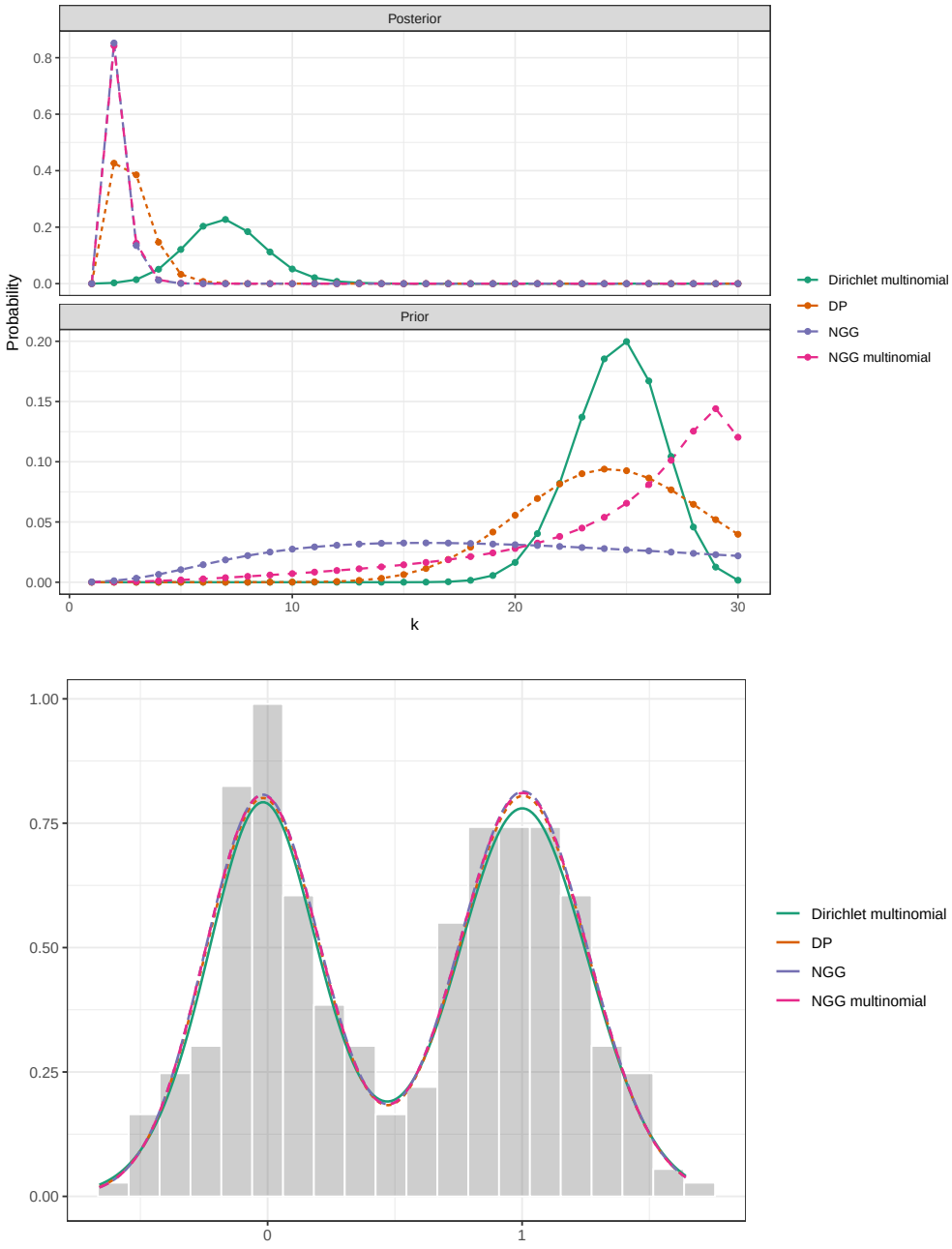

Figure 11: DATASET 2, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

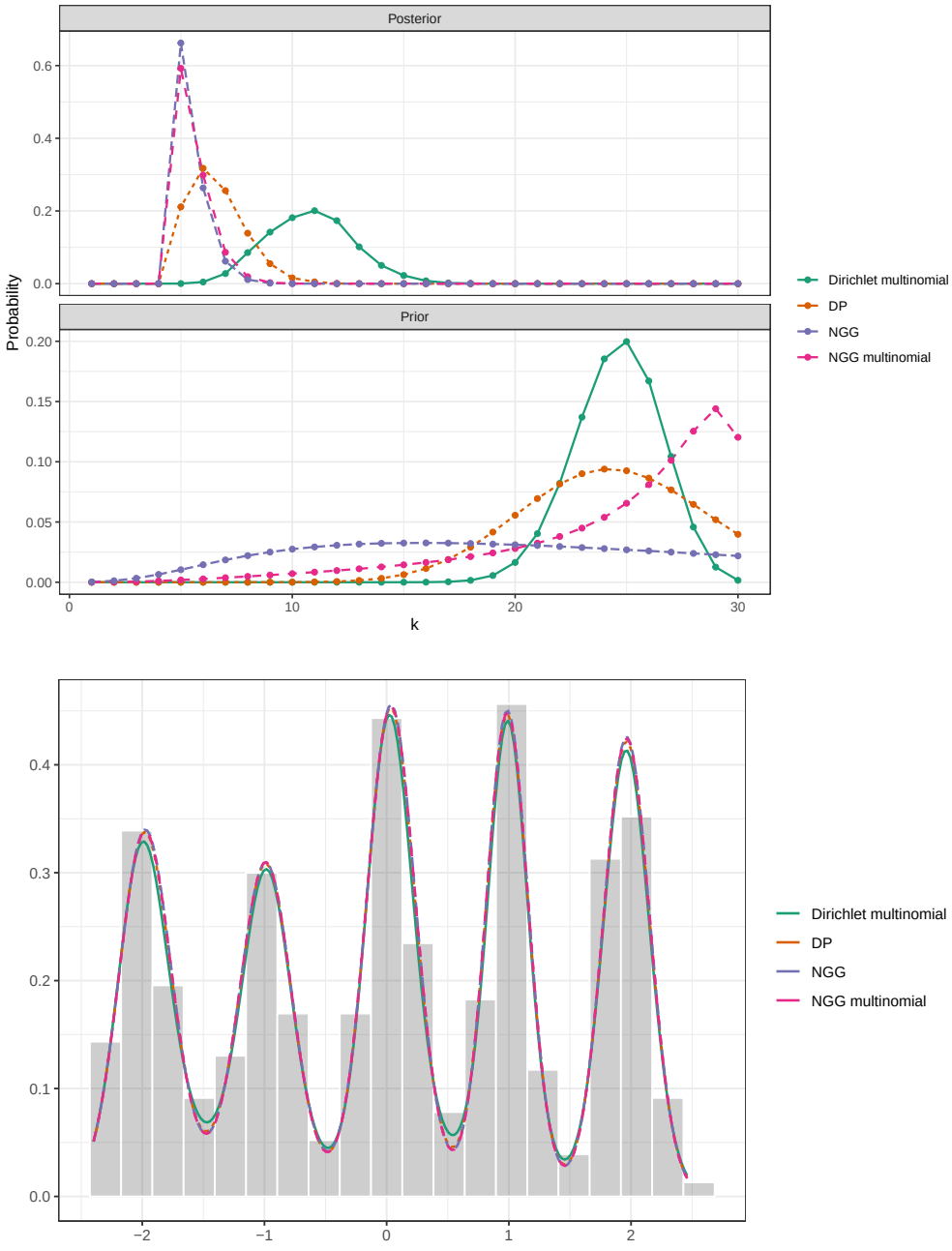

Figure 12: DATASET 3, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

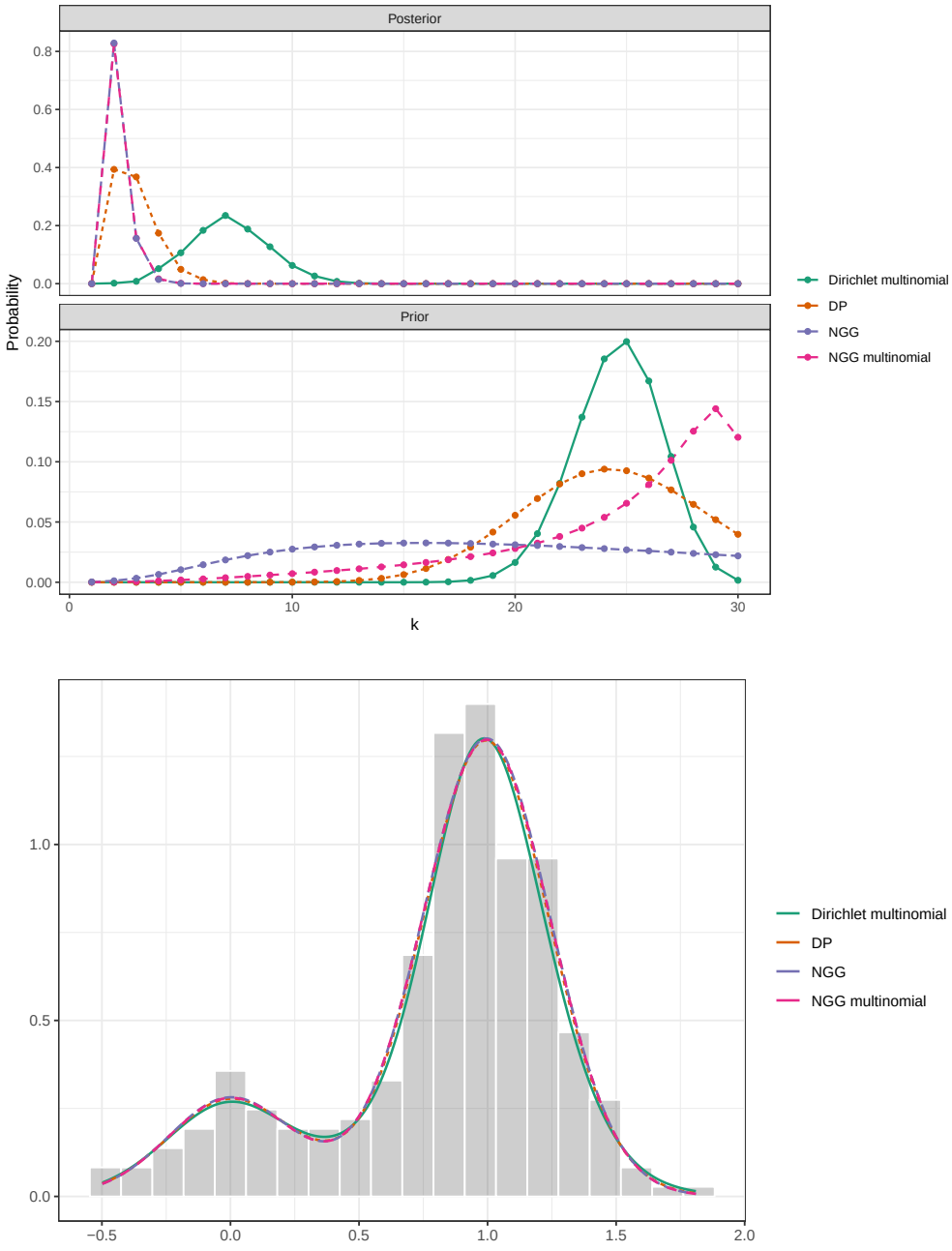

Figure 13: DATASET 4, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

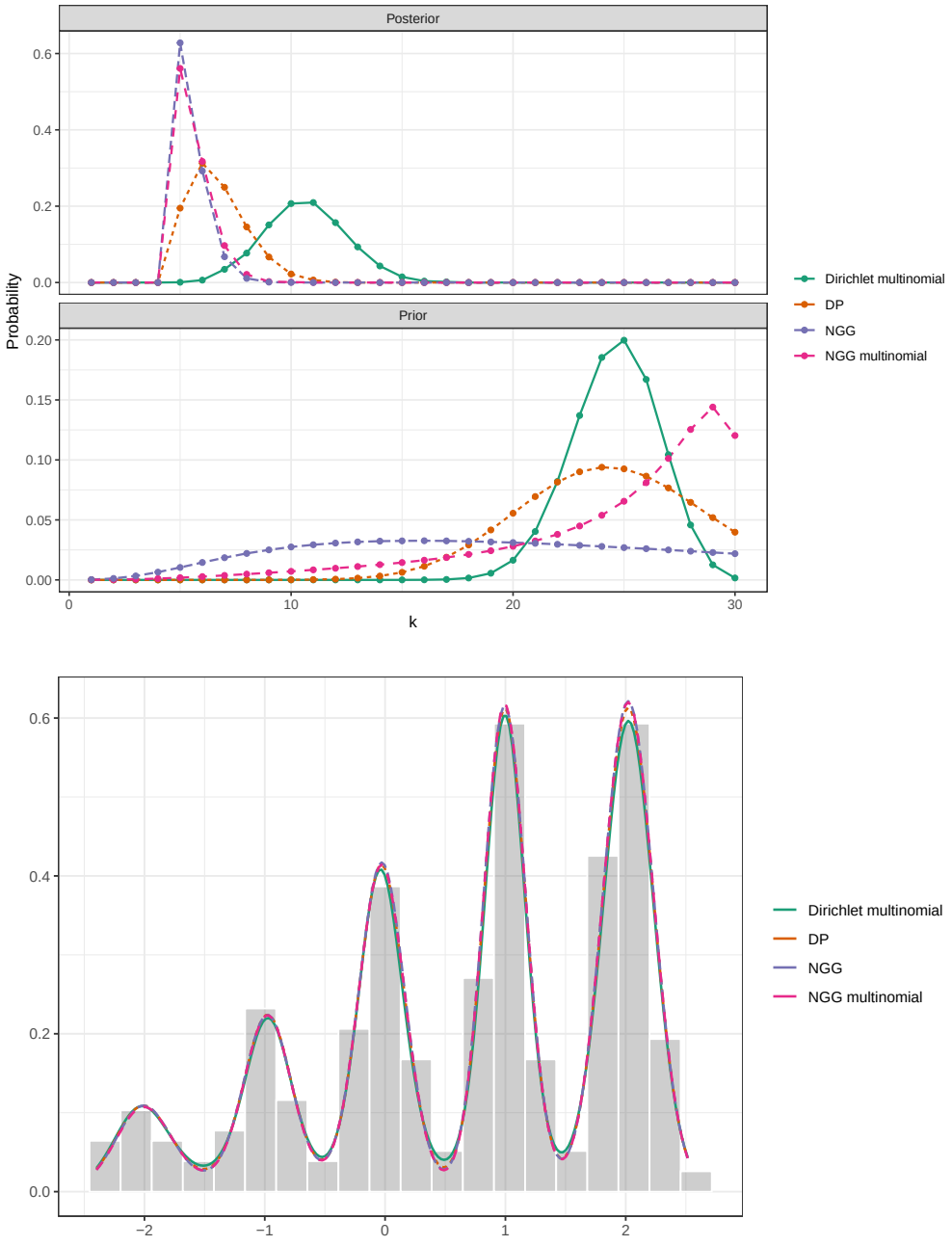

Figure 14: DATASET 5, SCENARIO B. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

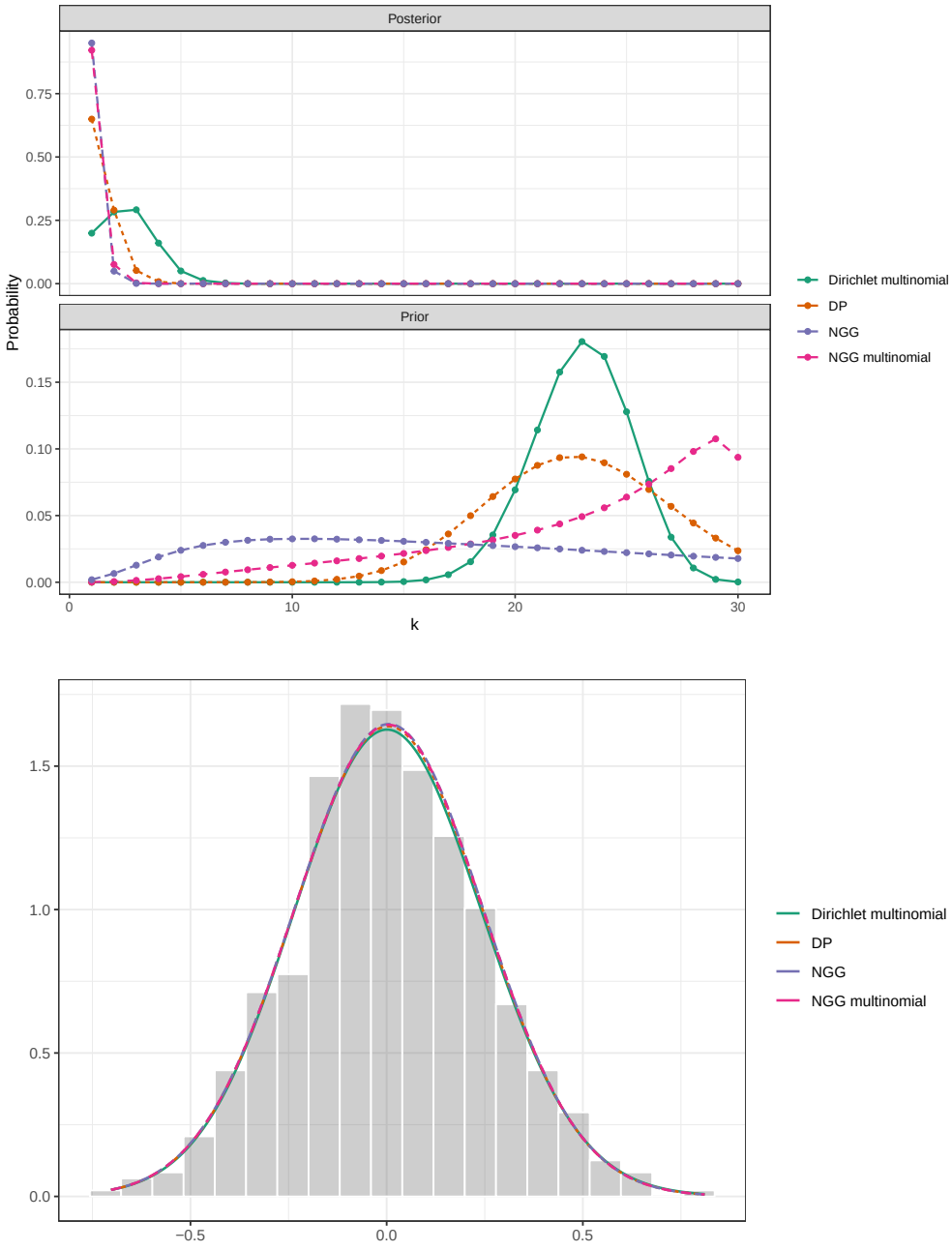

Figure 15: DATASET 1, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

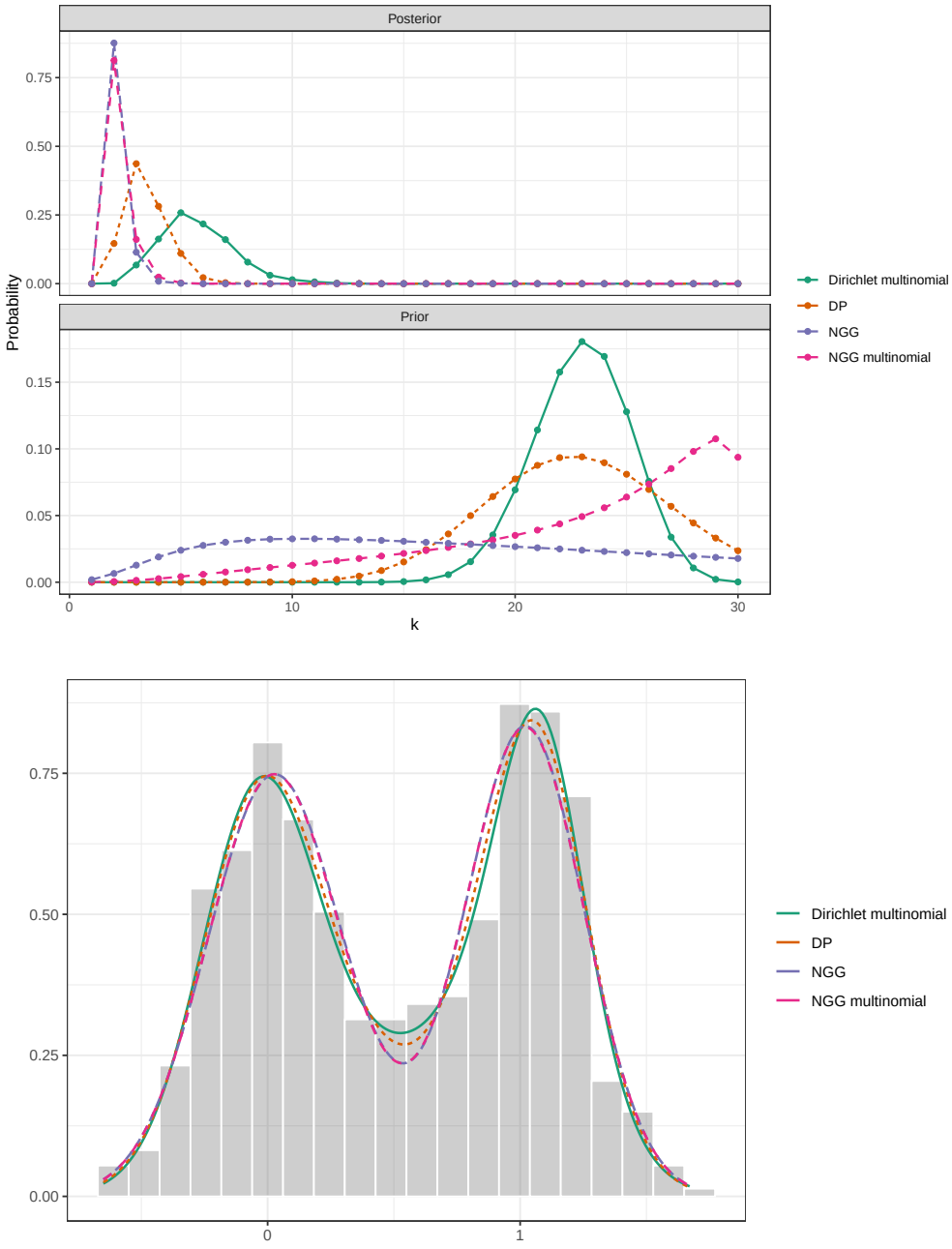

Figure 16: DATASET 2, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

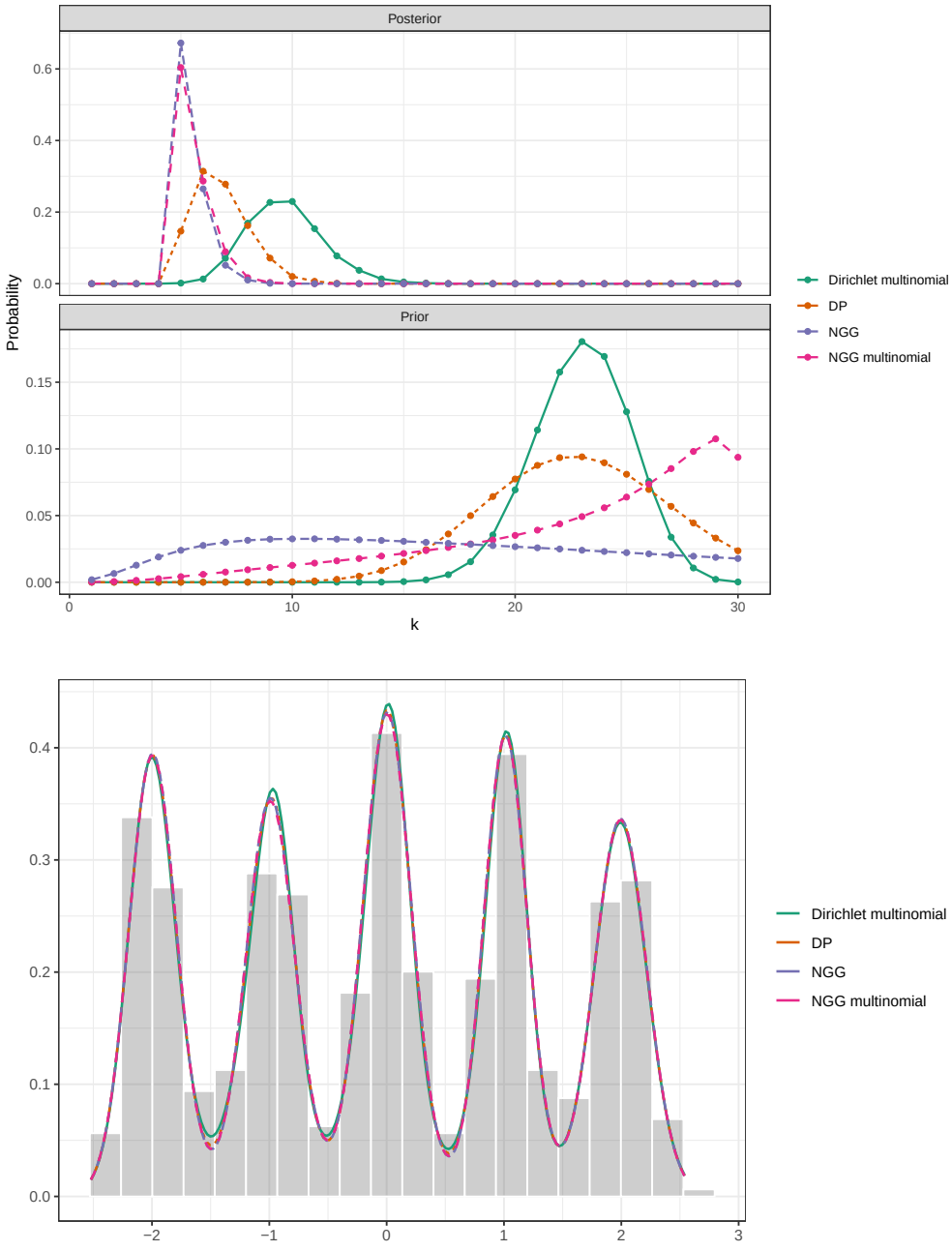

Figure 17: DATASET 3, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

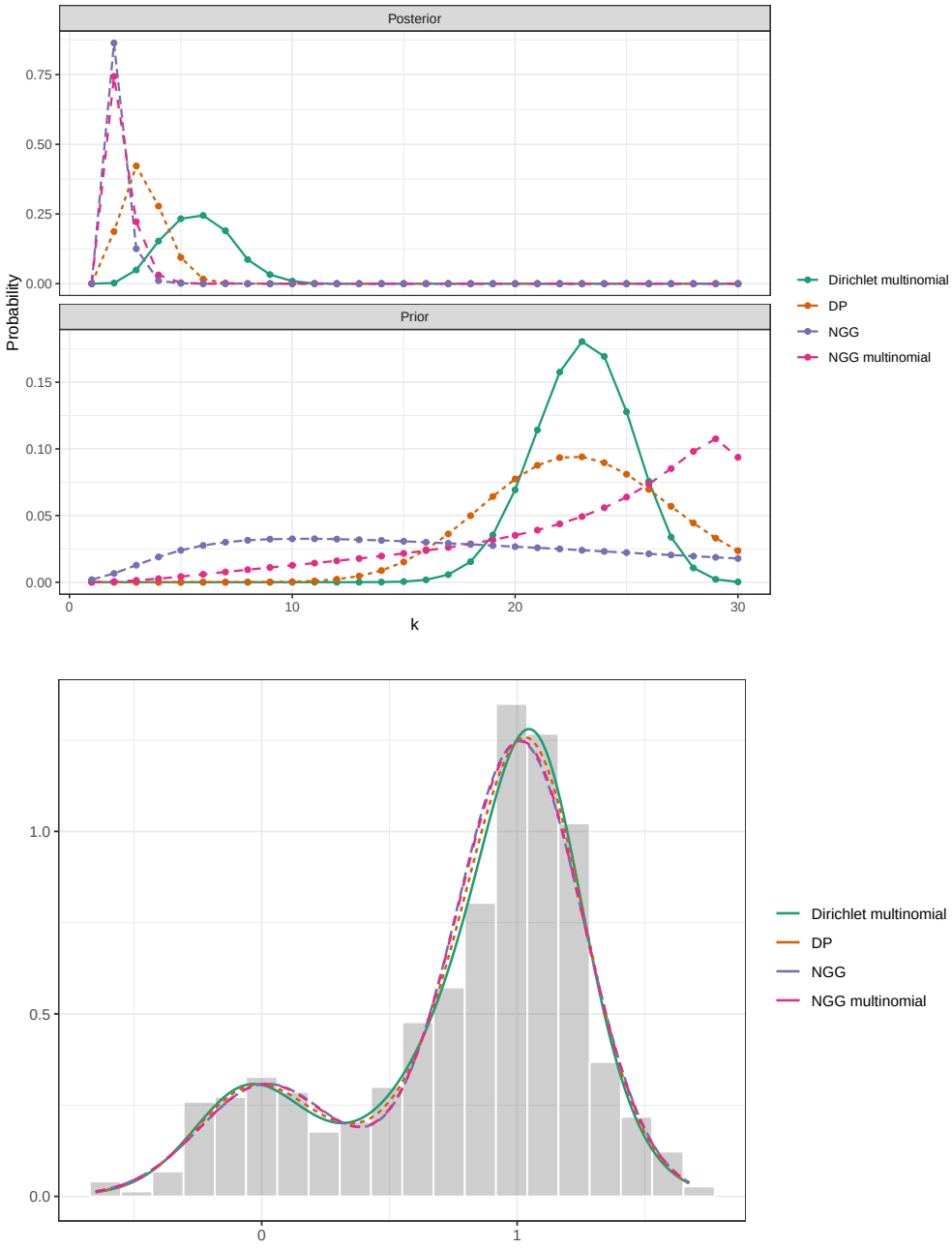

Figure 18: DATASET 4, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

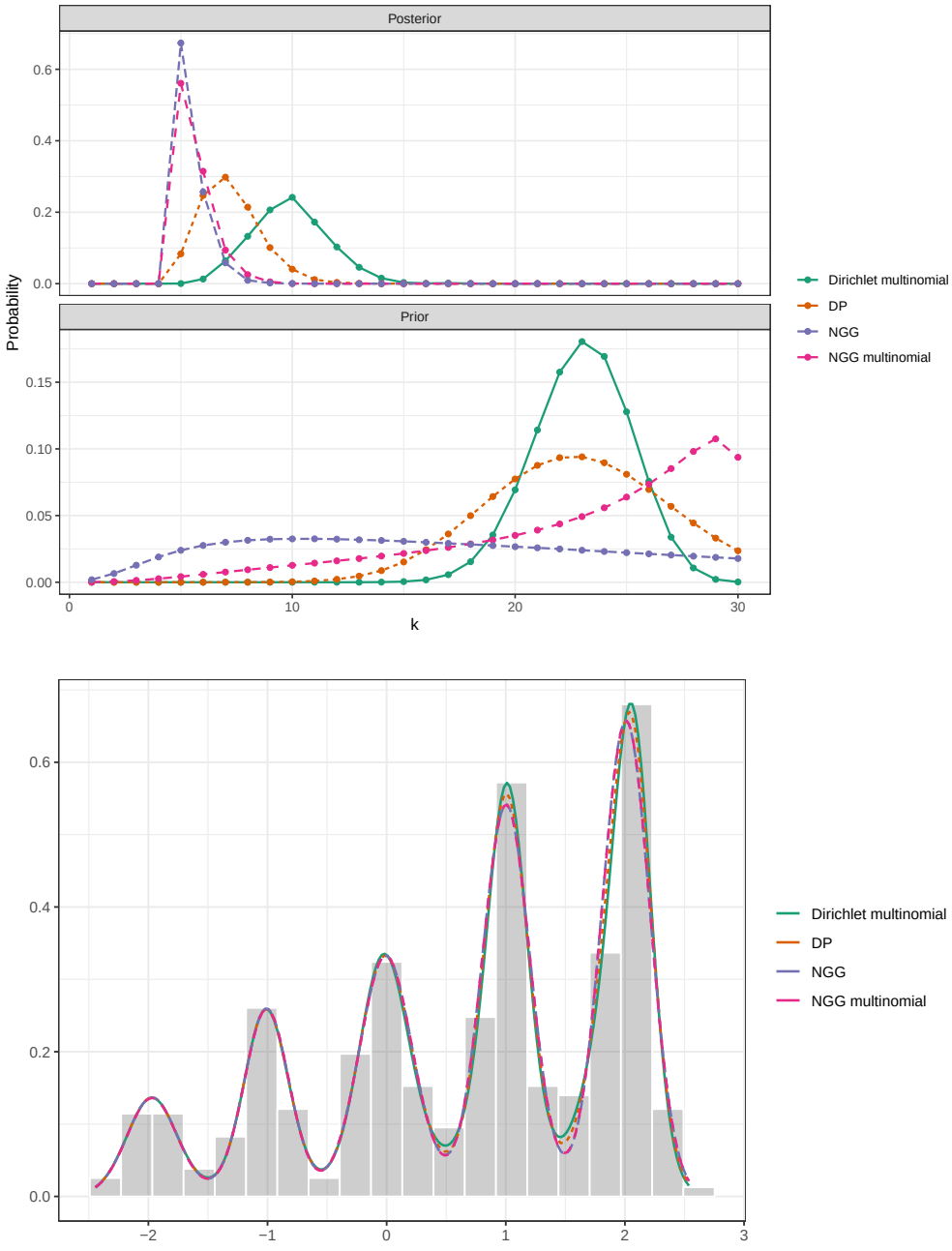

Figure 19: DATASET 5, SCENARIO C. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

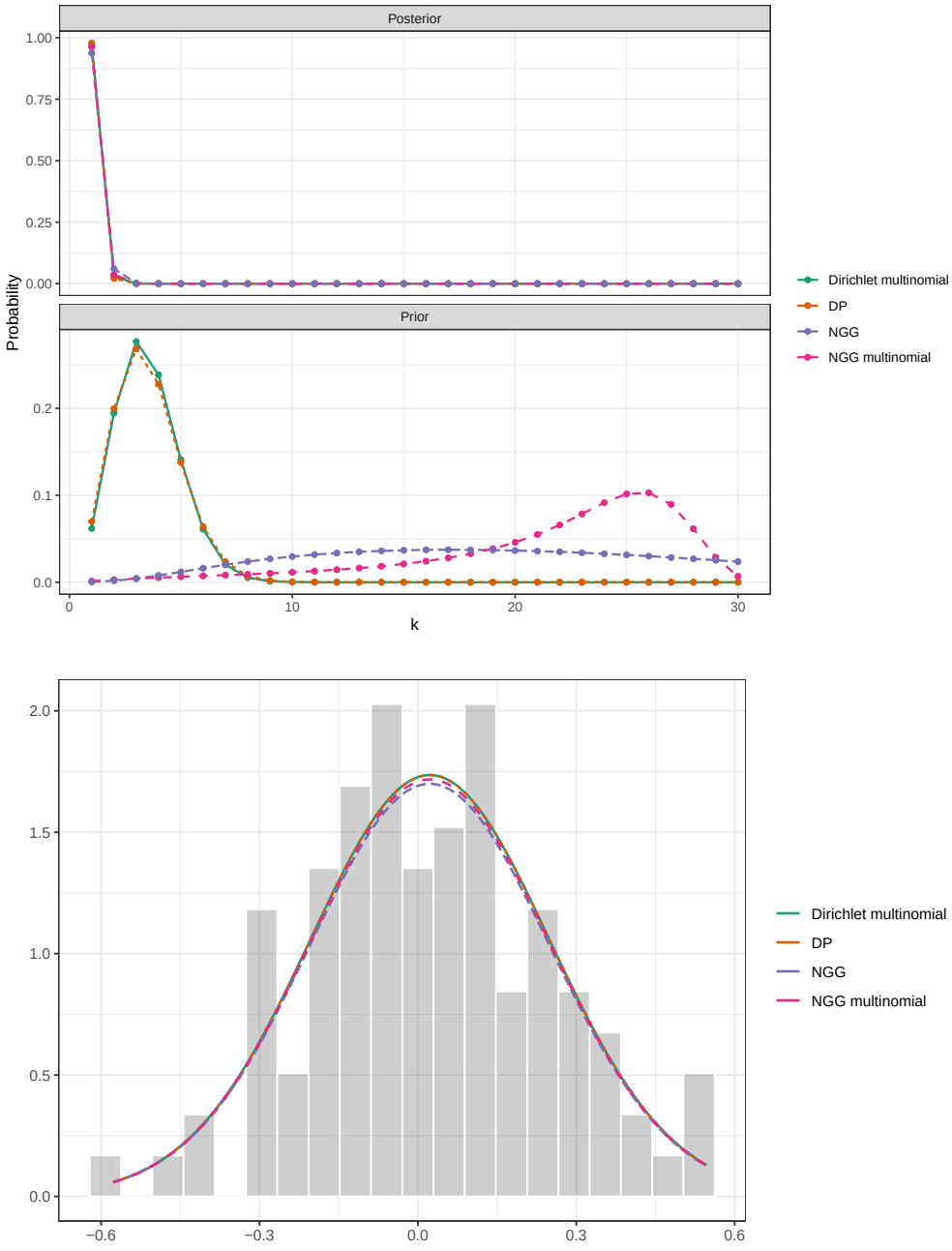

Figure 20: DATASET 1, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

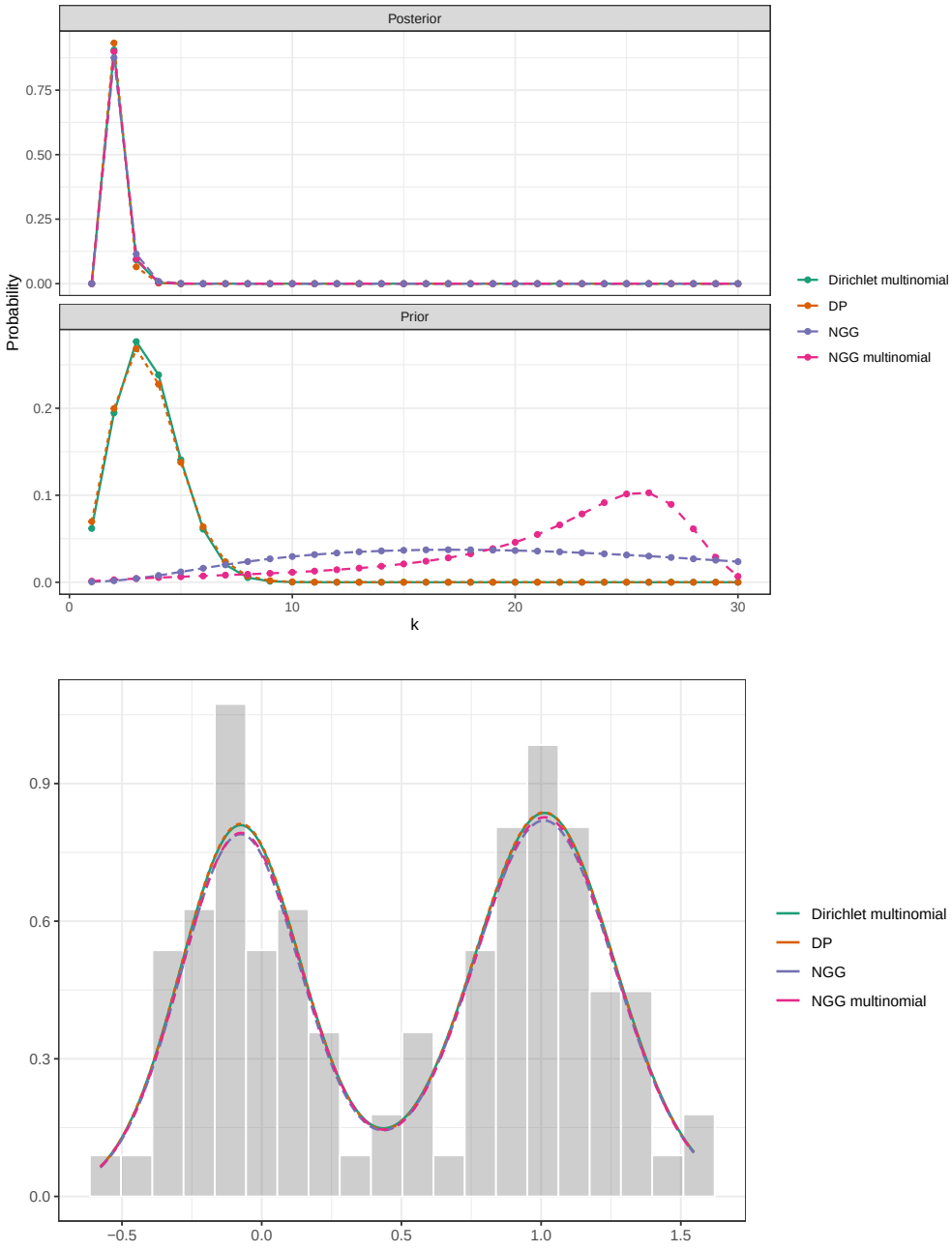

Figure 21: DATASET 2, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$ Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

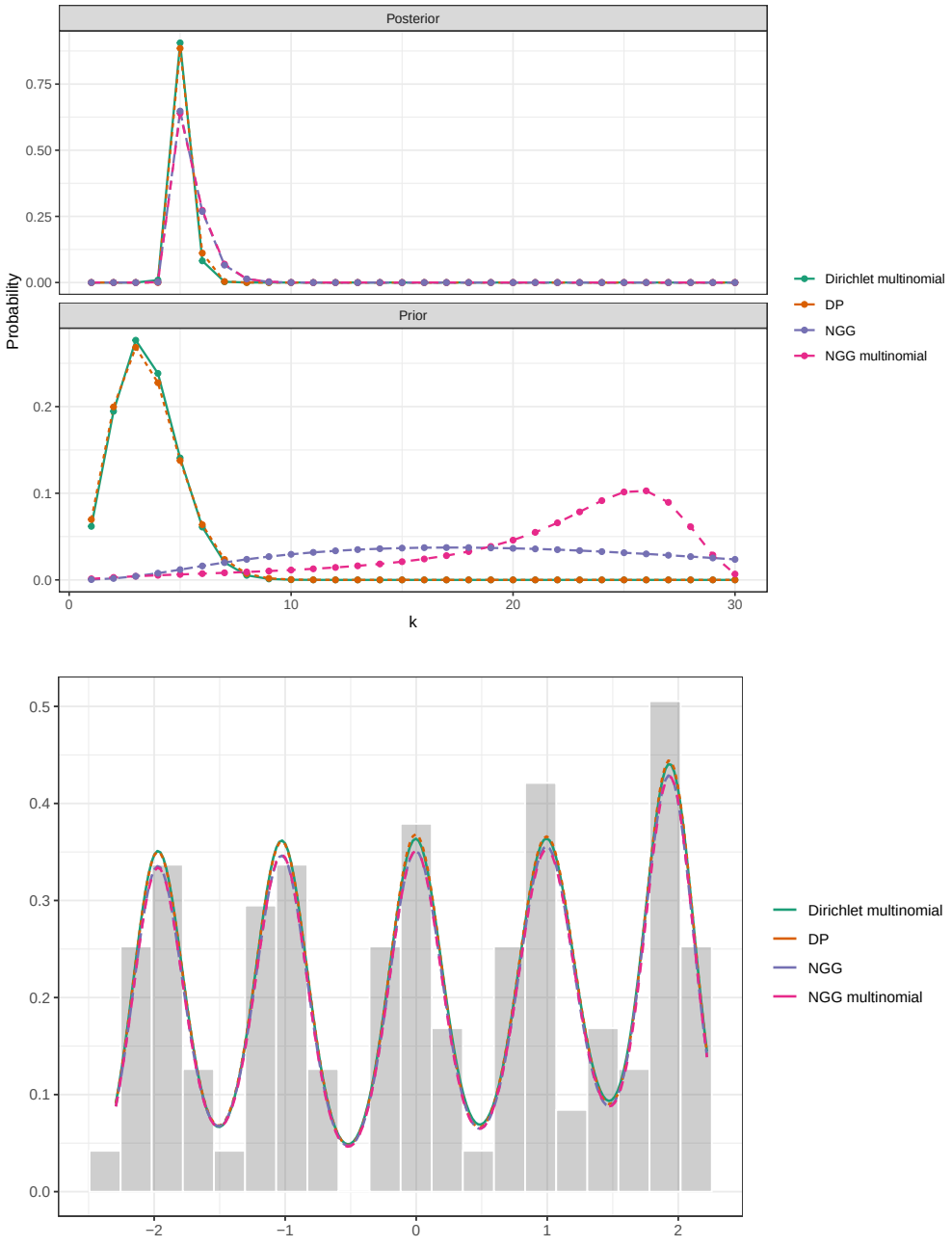

Figure 22: DATASET 3, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

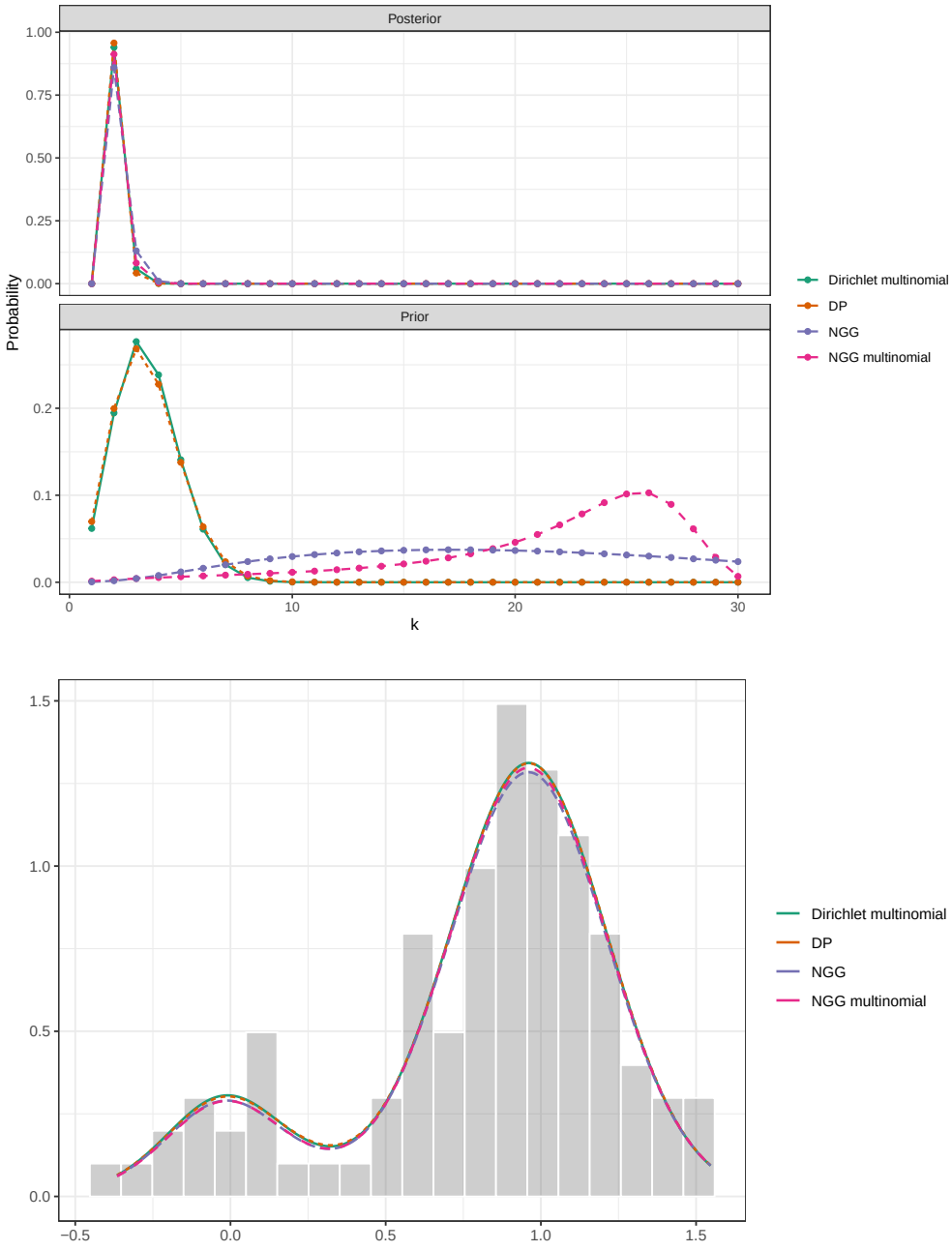

Figure 23: DATASET 4, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

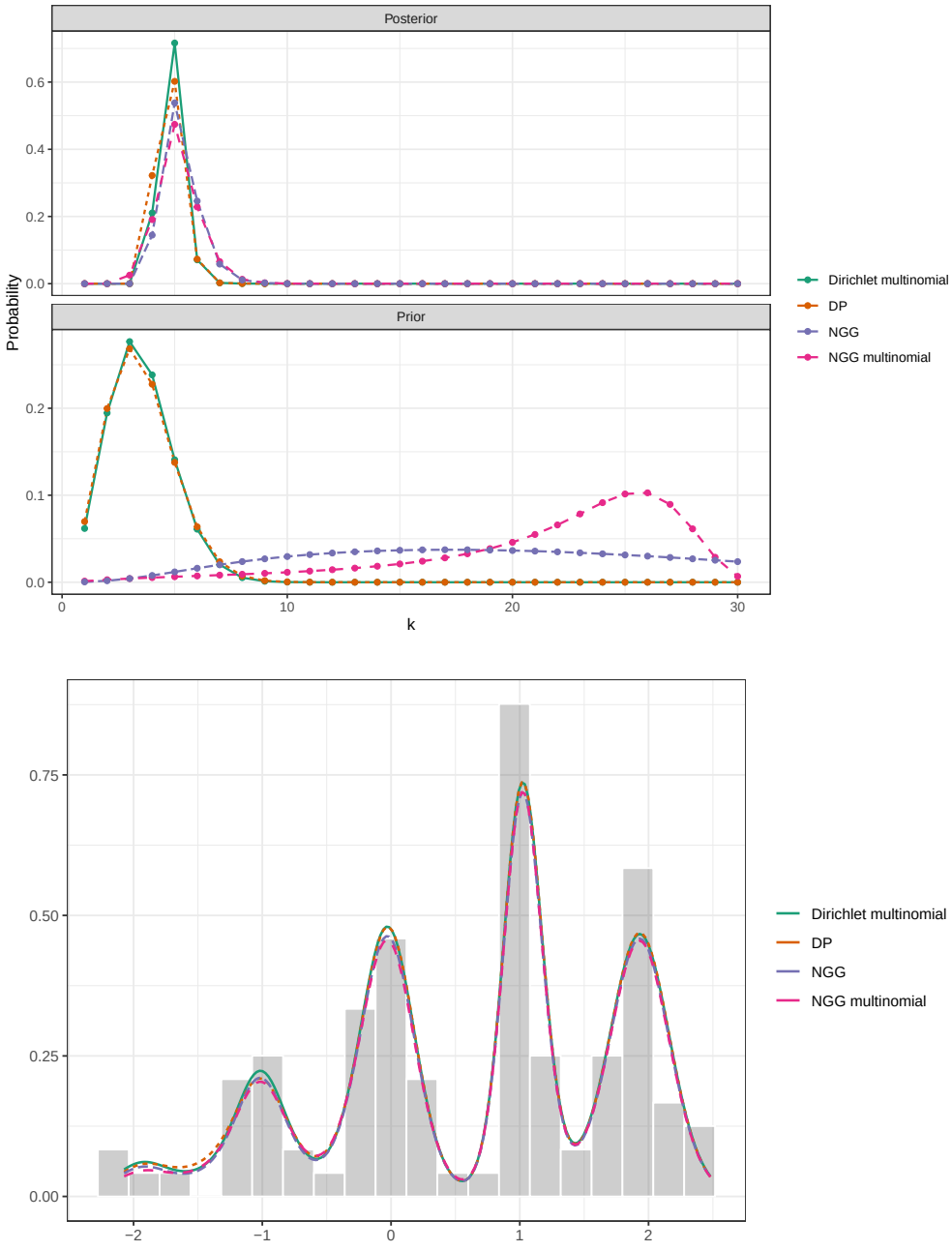

Figure 24: DATASET 5, SCENARIO D. Top panel: prior and posterior distributions of the number of clusters $K_{n,H}$. Bottom panel: lines represent the posterior mean of the density; gray bars correspond to the histogram of the raw data.

References

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.

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.

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

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

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.