Bayesian Analysis (0000) **00**, Number 0, pp. 1

Invited comment on Article by Camerlenghi, Dunson, Lijoi, Prünster and Rodríguez

Mario Beraha[*,†] and Alessandra Guglielmi[*]

We thank the authors (also denoted by Camerlenghi *et al.* hereafter) for a very interesting paper, which addresses the problem of testing homogeneity between two populations/groups. They start from pointing out a drawback of the Nested Dirichlet Process (NDP) by Rodriguez et al. (2008), i.e. its degeneracy to the exchangeable case: when the NDP is a prior for two population distributions (or for the corresponding mixing measures in mixture models), it forces homogeneity across the two samples in case of ties across samples at the observed or latent level. In fact, as pointed out by Camerlenghi *et al.*, the NDP does not accommodate for shared atoms across populations. This limitation, which is clear from the definition of NDP in Rodriguez et al. (2008), has a strong impact on the inference: as showed in this paper, if two populations share at least one common latent variable in the mixture model, the posterior distribution would either identify the two random measures associated to the populations as completely different (i.e. it would not recover the shared components) or it would identify them as identical. The need for a more flexible framework is elegantly addressed by the authors who propose a novel class of Latent Nested Nonparametric priors, where a shared random measure is added to the draws from a Nested Random Measure, hence accommodating for shared atoms. There are two key ideas in their model: (i) nesting discrete random probability measures as in the case of the nested Dirichlet process by Rodriguez et al. (2008), and (ii) contaminating the population distributions with a common component as in Müller et al. (2004) (or as in Lijoi et al., 2014). The latter yields dependency among the random probability measures of the populations and avoids the degeneracy issue pointed out by the authors, while the former accounts for testing homogeneity in two-sample problems.

As a comment on the computational perspective, we note that their MCMC method relies on the analytical expression of the Partially Exchangeable Partition Probability Function (pEPPF), which the authors obtain in the special case of $I = 2$ populations. However, the sampling scheme poses significant computational issues even in the case of $I = 2$, needing to rely on Monte Carlo integration to approximate some intractable integrals.

In this comment, we address the problem of extending their mixture model class for testing homogeneity of I populations, with $I > 2$, according to the first *path* the authors mention in their concluding remarks. In particular, we assume the mixture model for I populations/groups, when the mixing random probability measures $(\widetilde{p}_1, \ldots, \widetilde{p}_I)$ have a prior distribution that is the Latent Nested

[*]Politecnico di Milano, Milano, ITALY mario.beraha@polimi.it alessandra.guglielmi@polimi.it
[†]Also affiliated with Università degli Studi di Bologna

 DOI: 0000

Dirichlet process (LNDP) measure. This prior is more manageable than their general proposal, thanks to the stick-breaking representation of all the random probability measures involved, which can be easily truncated to give an approximation, which is straightforward to compute. Here, we apply the Latent Nested Dirichlet Process mixtures to simulated datasets from this paper, while the authors adopt a different latent nested nonparametric prior for $I = 2$ populations. By using the truncation approximation of stick breaking random probabilities, we do not need to resort to the pEPPF anymore and we are able to extend the analysis to cases with more than two populations.

However, our experience shows that this vanilla-truncation MCMC scheme does not scale well with I: the computational burden becomes demanding even for moderate values of I, which are common when testing homogeneity for different groups, for example while comparing a treatment in a small group of hospitals. If one assumes the LNDP as a prior for the mixing random probability measures $(\widetilde{p}_1, \ldots, \widetilde{p}_I)$, we have showed that we really need to derive either the posterior characterization of the LNDP, as suggested by the authors, or significantly more efficient truncation-based schemes.

1 Latent Nested Dirichlet Process mixture models

In this section, we make explicit the details of the definition of the Latent Nested Process that was introduced by the authors, and then consider the Latent Nested Dirichlet Process as the mixing distributions for I different populations. We also apply this model to synthetic data.

Consider the (Euclidean) space Θ and let $\mathbb{M}_\Theta$ be the space of all bounded measures on Θ. Let $\tilde{q}$ be a random probability measure, $\tilde{q} \sim \text{NRMI}[\nu, \mathbb{M}_\Theta]$ with intensity $\nu(ds, dm) = c\rho(s)dsQ(dm)$; here $c > 0$, ρ is a function defined on $\mathbb{R}^+$ under conditions

$$\int_0^{+\infty} \min\{1, s\}\rho(s)ds < +\infty, \quad \int_0^{+\infty} \rho(s)ds = +\infty,$$

and Q is a probability measure on $\mathbb{M}_\Theta$. We skip the details on the σ-algebras attached to the spaces we consider. We know that $\tilde{q} = \sum_{j=1}^{\infty} \tilde{\omega}_j \delta_{\tilde{\eta}_j}$, where $\{(\tilde{\omega}_j, \tilde{\eta}_j)\}$ are the points of a Poisson process with mean intensity $\nu(ds, dm)$. In particular, $\tilde{\eta}_j \overset{\text{iid}}{\sim} Q$, i.e. each $\tilde{\eta}_j$ is itself a CRM on Θ with Lévy intensity $\nu_0(ds, d\theta) = c_0\rho_0(s)dsQ_0(d\theta)$, which implies $\tilde{\eta}_j = \sum_{k=1}^{\infty} J_k^j \delta_{\theta_k^j}$, where, for each j, $\{(J_k^j, \theta_k^j), k \geq 1\}$ are the points of a Poisson process with mean intensity $\nu_0(ds, d\theta)$. Here $c_0 > 0$, ρ_0 is a function on $\mathbb{R}^+$ under the same conditions as $\rho(s)$ and Q_0 is a probability measure on Θ. Finally, let q_S be the law of μ_S, a CRM on Θ, with Lévy intensity $\nu_0^* = \gamma\nu_0$, where $\gamma > 0$.

Similarly to the authors, we define a Latent Nested Process as a collection of random probability measures $\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_I$ on Θ such that

$$\tilde{p}_i = \frac{\mu_i + \mu_s}{\mu_i(\mathbb{X}) + \mu_s(\mathbb{X})} = w_i \frac{\mu_i}{\mu_i(\mathbb{X})} + (1 - w_i)\frac{\mu_S}{\mu_S(\mathbb{X})}, \; i = 1, \ldots, I,$$

where

$$\mu_1, \mu_2, \ldots, \mu_I, \mu_S | \tilde{q}, q_S \sim \tilde{q} \times \tilde{q} \ldots \times \tilde{q} \times q_S.$$

In particular, if we set $\rho(s) = \rho_0(s) = s^{-1}\mathrm{e}^{-s}$, $s > 0$, we obtain the Latent Nested Dirichlet Process; since the μ_i's and μ_S are independent gamma processes in this case, the μ_i's also being iid, and

$$p_i = \frac{\mu_i}{\mu_i(\mathbb{X})}, i = 1, \ldots, I, \quad p_S = \frac{\mu_S}{\mu_S(\mathbb{X})},$$

i.e. p_i and p_S are draws from two independent Dirichlet processes, we have

$$p_i \mid G \overset{\text{iid}}{\sim} G = \sum_{l=1}^{\infty} \pi_l \delta_{G_l^*}, \ i = 1, \ldots, I, \tag{1.1}$$

$$p_S = \sum_{h=1}^{\infty} w_h^S \delta_{\theta_h^S}, \tag{1.2}$$

where G is a Nested Dirichlet process, i.e. a DP whose atoms are DPs. We use notation $(\tilde{p}_1, \ldots, \tilde{p}_I) \sim LNDP(\gamma, \nu_0, \nu)$ for

$$\widetilde{p}_i = w_i p_i + (1 - w_i) p_S, \quad i = 1, \ldots, I.$$

Note that each $\widetilde{p}_i$ is a mixture of two components: an idiosyncratic component p_i and a shared component p_S, where the latter preserves heterogeneity across populations even when shared values are present. As pointed out by the authors, the random indicator functions of the two events $\widetilde{p}_i = \widetilde{p}_{i'}$ and $p_i = p_{i'}$ coincide a.s., if $i \neq i'$. This latter event has positive prior probability for any couple of distinct indexes i, i' in $\{1, \ldots, I\}$. Summing up, this prior induces a prior distribution for the parameter $\boldsymbol{\rho}$, the partition of population indexes $\{1, 2, \ldots, I\}$: two populations are clustered together if they share the same mixing measure.

Now, suppose that we have data from I different populations (e.g. measurements on patients in different hospitals). Let y_{ji}, $j = 1, \ldots, n_i$, be observations for different subjects in population i, for $i = 1, \ldots, I$. We assume that, for any $i = 1, \ldots, I$,

$$\begin{aligned} y_{ji} \mid \widetilde{p}_i &\overset{\text{iid}}{\sim} \int_{\Theta} f(y_{ji} \mid \theta) \widetilde{p}_i(d\theta), \quad j = 1, \ldots, n_i \\ (\widetilde{p}_1, \ldots \widetilde{p}_I) &\sim LNDP(\gamma, \nu_0, \nu). \end{aligned} \tag{1.3}$$

For computing posterior inference, instead of considering model (1.3), we consider a truncation approximation of the stick-breaking representation of the LNDP, similarly as in Rodriguez et al. (2008). In particular, instead of (1.1)-(1.2), we consider the p_i's iid from a L-H truncation of a nested Dirichlet process, i.e.,

$$\begin{aligned} p_i | G &\overset{\text{iid}}{\sim} \sum_{l=1}^{L} \pi_l \delta_{G_l^*}, \qquad \pi_l = \nu_l \prod_{s=1}^{l-1} (1 - \nu_s), \quad \nu_l \overset{\text{iid}}{\sim} Beta(1, c) \quad l = 1, \ldots, L-1, \quad \nu_L = 1 \\ G_l^* &= \sum_{h=1}^{H} w_{lh} \delta_{\theta_{lh}^*}, \quad w_{lh} = u_{lh} \prod_{s=1}^{h-1} (1 - u_{ls}), \quad u_{lh} \overset{\text{iid}}{\sim} Beta(1, c_0) \quad h = 1, \ldots, H-1, \quad u_{lH} = 1 \\ &\theta_{lh}^* \overset{\text{iid}}{\sim} Q_0 \text{ for all } l, h \end{aligned}$$

and p_S itself is an H–truncated Dirichlet Process of parameters γc_0 and Q_0. Since w_i is defined from the total masses of independent gamma processes, then

$$w_i = \frac{\mu_i(\Theta)}{\mu_i(\Theta) + \mu_S(\Theta)} \sim Beta(c_0, \gamma c_0), \quad i = 1, \ldots, I.$$

This truncation approximation could be exploited to design blocked Gibbs sampling schemes as in Ishwaran and James (2001), or more general truncation schemes (see the references in Argiento et al., 2016); in the next section we use this truncation approximation in order to write a JAGS code to fit the data from the examples.

2 Simulation Study

We have fitted the truncated Latent Nested Dirichlet Process mixture model to simulated data via JAGS, using $L = 30$ and $H = 50$. The parametric kernel $f(y|\theta)$ in (1.3) is the unidimensional Gaussian density with mean θ and variance σ^2, i.e. $\theta = (\mu, \sigma)$. For every simulated dataset, we have considered the base measure $Q_0(\mu, \sigma) = \mathcal{N}\left(0, \lambda\sigma^2\right) \times \mathcal{U}(\sigma \mid 0, 2)$, with $\lambda = 10$. Moreover we set $c = c_0 = 1$ and let $\gamma \sim \mathcal{U}(0.25, 5)$. Chains were run for 10,000 iterations after 15,000 iterations of adaptation and 5,000 iterations of burn-in, thinning every 10 iterations for a final sample size equal to 1,000.

First, we considered two of the simulated scenarios examined in the paper, specifically scenarios I and II, and we simulated $n_1 = n_2 = 100$ observations from each group. Scenario I corresponds to full exchangeability across two groups of data, i.e.

$$y_{j1}, y_{j2} \overset{\text{iid}}{\sim} 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(5, 1),$$

while scenario II corresponds to partial exchangeability with a shared component between the populations

$$y_{j1} \overset{\text{iid}}{\sim} 0.9\mathcal{N}(5, 0.6) + 0.1\mathcal{N}(10, 0.6) \quad y_{j2} \overset{\text{iid}}{\sim} 0.1\mathcal{N}(5, 0.6) + 0.9\mathcal{N}(0, 0.6).$$

Both scenarios were tested in the paper under the same Gaussian kernel we consider, with a latent nested σ-stable mixture model instead of the LNDP as a prior for the mixing distributions. We have considered another simulated dataset from $I = 3$ populations, with $n_1 = n_2 = n_3 = 100$, that is

$$y_{j1} \overset{\text{iid}}{\sim} 0.2\mathcal{N}(5, 0.6) + 0.8\mathcal{N}(0, 0.6) \quad y_{j2} \overset{\text{iid}}{\sim} 0.2\mathcal{N}(5, 0.6) + 0.8\mathcal{N}(0, 0.6) \quad y_{j3} \overset{\text{iid}}{\sim} \mathcal{N}(-3, 0.6),$$

which corresponds to full exchangeability across populations $1, 2$ but not across $1, 2, 3$.

As pointed out by the authors, Bayes factors for homogeneity tests across populations are available as a by-product of their model. Homogeneity tests with hypotheses

$$H_0 : \tilde{p}_i = \tilde{p}_j \quad vs \quad H_1 : \tilde{p}_i \neq \tilde{p}_j \tag{2.1}$$

are performed by the authors in case $(i, j) = (1, 2)$, by introducing the auxiliary variable $\mathbb{I}_{\{\tilde{p}_1 = \tilde{p}_2\}}$ in their MCMC state space, so that draws from its posterior are straightforwardly available. In our formulation of the LNDP mixture model instead, we resort to the cluster allocation variables of the nested process, $s_j = l$ iff $p_j = G_l^*$ for $j = 1, \ldots, I$, to perform the same tests.

In case of $I > 2$ populations, it is also possible to perform global tests on the cluster structure arising among the populations. In our new (third) scenario, we are interested in testing the presence of one single group against the presence of three groups (for example), i.e.

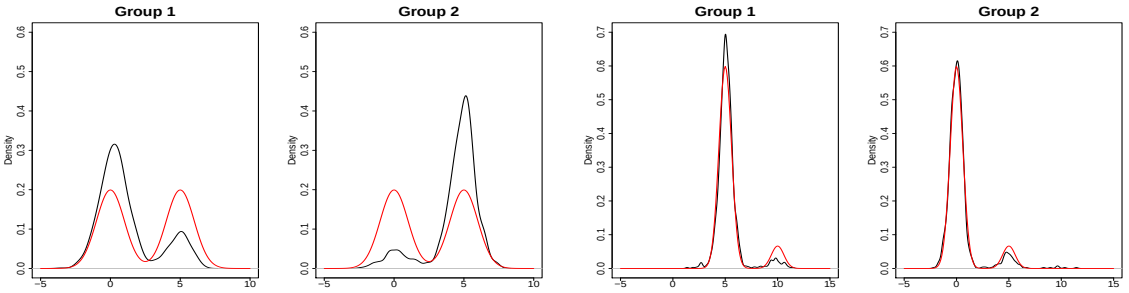
$$H_0 : \tilde{p}_1 = \tilde{p}_2 = \tilde{p}_3 \quad vs \quad H_1 : \tilde{p}_1 \neq \tilde{p}_2 \neq \tilde{p}_3.$$

This type of tests are straightforward to obtain, since they are based on the EPPF of the nested process. Indeed, a priori, $P(\tilde{p}_1 = \tilde{p}_2 = \tilde{p}_3) = P(\boldsymbol{\rho} = \{1, 2, 3\})$ while $P(\tilde{p}_1 \neq \tilde{p}_2 \neq \tilde{p}_3) = P(\boldsymbol{\rho} = \{1\}, \{2\}, \{3\})$, where $\boldsymbol{\rho}$ is the partition of $\{1, 2, 3\}$ arising from the nested process; posterior odds are obtained once again monitoring the values of the allocation variables s_j's. The Bayes factor for this specific test equals 0.18, providing evidence in favour of H_1.

Scenario	(i, j)	BF_{01}
I	(1, 2)	1.00
II	(1, 2)	0.08
3 populations	(1, 2)	1.27
	(1, 3)	0.07
	(2, 3)	0.09

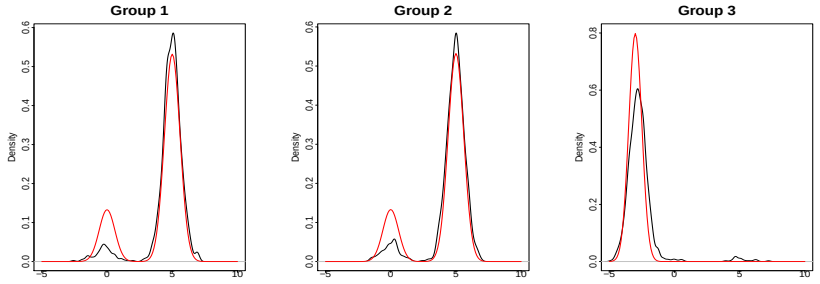
Table 1: Bayes factors for hypotheses (2.1) under the three simulated scenarios.

Table 1 reports the Bayes factors for tests (2.1) computed via our MCMC, while Figure 1 displays the predictive densities in each population. As far as the Bayes factors are concerned, we have computed those corresponding to hypotheses (2.1) with $(i, j) = (1, 2)$ for scenarios I and II, while for the new scenario we consider all the possible pairwise tests, i.e. $(i, j) = (1, 2),\ (1, 3),\ (2, 3)$. The Bayes factors in Table 1 correctly indicate strong evidence in favour of the alternative hypothesis for the second and third test of the 3-populations scenario, as well as for scenario II, while for the other tests there is no clear evidence in either direction. The BF_{01} for scenario II is much larger than the corresponding Bayes factor computed by the authors, obtained under the latent nested σ-stable mixture model; similarly, our BF_{01} for scenario I is equal to 1, while the authors obtain a larger value, giving evidence in favour of the true hypothesis. Of course, the mixing of the chain produced by JAGS, especially for scenario I with equal mixture weights, is generally worse than any specifically-designed MCMC scheme, as the one described by the authors. However, the density estimates (in black) for scenario II in Figure 1(b) are accurate, unlike those in Figure 1(a) where we clearly see that the JAGS code is not able to recover the weights in the true density in each group, while recovering the locations. Predictive densities in Figure 1(c) are close to the true population distributions in all the groups, even though we experienced the same difficulties in recovering the



(a) Scenario I

(b) Scenario II



(c) New scenario

Figure 1: Density estimates for scenario I (a), II (b) and the new scenario with $I = 3$ populations (c). In every panel, the black line denotes the predictive density in the population, while the red line is the density which generated the data.

true weights of all the mixtures because of the large number of allocation parameters in the JAGS model, which makes sampling much less efficient.

To conclude our experiments, we have also designed a scenario with 4 populations simulating $n_i = 100$ observations from each true population distribution, which is a mixture of two Gaussian components. The Bayes factors for hypotheses (2.1), computed via our JAGS MCMC, are in agreement with the true underlying clustering, that is $\{1, 2\}, \{3, 4\}$. However, even with as little as 100 observations per group, the MCMC simulation took more than 8 hours to run. To make a comparison, in our experience, the runtime of our JAGS code for $I = 3$ populations was about 2.5 times longer than for $I = 2$ populations, and that for $I = 4$ groups was approximately 4 times larger than for $I = 2$.

Despite the construction of ad-hoc Gibbs sampling schemes, possibly based on the truncated stick breaking representation, which could greatly improve the performances we reported, we believe that this model, generalized as we have presented here to the case of $I > 2$ populations and using a truncation approximation for the LNDP, contains inherent computational difficulties which are not easy to deal with. Assuming a larger value for I, even though a moderate value as in case of, e.g., comparing a patient treatment in a few dozens of hospitals, will still be challenging using the model we have considered here, taking into action the suggestion Camerlenghi *et al.* made in their concluding remarks.

References

Argiento, R., Bianchini, I., and Guglielmi, A. (2016). "A blocked Gibbs sampler for NGG-mixture models via a priori truncation." *Statistics and Computing*, 26(3): 641–661.

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173.

Lijoi, A., Nipoti, B., Prünster, I., et al. (2014). "Bayesian inference with dependent normalized completely random measures." *Bernoulli*, 20(3): 1260–1291.

Müller, P., Quintana, F., and Rosner, G. (2004). "A method for combining inference across related nonparametric Bayesian models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3): 735–749.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association*, 103(483): 1131–1154.