Contributed Discussion

Emanuele Aliverti[*], Sally Paganin[†], Tommaso Rigon[‡], and Massimiliano Russo[§,¶]

We congratulate the authors on an interesting paper, which provides a concrete contribution in Bayesian nonparametric methods. The proposed latent nested process (LNP) of Camerlenghi *et al.* is a notable generalization of the nested Dirichlet process (NDP) of Rodríguez et al. (2008). In the first place, Camerlenghi *et al.* extend the NDP to a broader class of nested processes (NP), leveraging on homogeneous random measures with independent increments (Regazzini et al., 2003). They elegantly frame this novel class of priors within the theory of completely random measures.

The rigorous theoretical study of the involved clustering mechanism allows Camerlenghi *et al.* to identify a potential pitfall of general NPs. Specifically, two random discrete distributions $\tilde{p}_\ell$ and $\tilde{p}_{\ell'}$, associated to different groups (populations) and distributed according to a NP, are either identical (i.e. $\tilde{p}_\ell = \tilde{p}_{\ell'}$ a.s.), or they do not have common atoms. This behavior implies that NPs can borrow information across groups only in an extreme fashion, that is, by assuming full homogeneity across populations. In contrast, the LNP generalization accommodates smooth transitions between the full homogeneity and the independence cases, while still accounting for clustering across different populations.

We will focus on the latent nested Dirichlet process special case, which has been considered by Camerlenghi *et al.* in their Example 2. First recall that the NDP of Rodríguez et al. (2008), in presence of $d \geq 2$ populations, can be alternatively defined through a Blackwell and MacQueen (1973) urn-scheme. Let δ_x denote a point mass at x. If $(p_1, \ldots, p_d)$ is a collection of random probability measures on a complete and separable metric space $\mathbb{X}$ following an NDP, then for any $c > 0$

$$p_{\ell+1} \mid p_1, \ldots, p_\ell \sim \frac{c}{c+\ell} Q + \frac{1}{c+\ell} \sum_{i=1}^{\ell} \delta_{p_i}, \qquad \ell = 1, \ldots, d-1, \tag{1}$$

where Q is the probability distribution of a Dirichlet process $\tilde{q}_0 \sim \text{DP}(c_0 Q_0)$, with precision parameter $c_0 > 0$ and with Q_0 being a non-atomic probability measure on $\mathbb{X}$. In other words, each p_ℓ is either a sample from a $\text{DP}(c_0 Q_0)$ or is set equal to one of the previously observed random measures.

The latent nested Dirichlet process is built upon (1). More precisely, the vector of random probability measures $(\tilde{p}_1, \ldots, \tilde{p}_d)$ characterizing such a process is obtained as a

[*]Department of Statistical Sciences, Università degli studi di Padova, Padova, Italy, aliverti@stat.unipd.it

[†]Department of Environmental Science, Policy & Management, University of California Berkeley, Berkeley, USA, sally.paganin@berkeley.edu

[‡]Department of Statistical Science, Duke University, Durham, USA, tommaso.rigon@duke.edu

[§]Harvard–MIT Center for Regulatory Science, Harvard Medical School, Boston, USA, massimiliano_russo@hms.harvard.edu

[¶]Department of Data Science Dana-Farber Cancer Institute, Boston, USA

convex combination of two random probability measures, namely

$$\tilde{p}_\ell = w_\ell p_\ell + (1 - w_\ell) p_S, \qquad \ell = 1, \dots, d, \tag{2}$$

where $p_S \sim \text{DP}(\gamma c_0 Q_0)$, with $\gamma > 0$, is independent on $p_1, \dots, p_d$ whereas $w_\ell \overset{\text{iid}}{\sim} \text{Beta}(c_0, \gamma c_0)$, independently on the random probability measures $p_1, \dots, p_d$ and p_S.

As formalized by Proposition 4 in Camerlenghi *et al.*, some random probability measures among $\tilde{p}_1, \dots, \tilde{p}_d$ will be identical with positive probability. Broadly speaking, this occurs if ties are present in the underlying urn-scheme of (1). In the Dirichlet case, the *a priori* probability of homogeneity among two distributions is

$$\pi_1^* := \mathbb{P}(p_\ell = p_{\ell'}) = \mathbb{P}(\tilde{p}_\ell = \tilde{p}_{\ell'}) = \frac{1}{c+1}, \qquad \ell \neq \ell'.$$

Thus, Camerlenghi *et al.* suggest to evaluate the posterior probability $\mathbb{P}(p_\ell = p_{\ell'} \mid \boldsymbol{X})$, to test the null hypothesis $H_0 : \tilde{p}_\ell = \tilde{p}_{\ell'}$ against the alternative $H_1 : \tilde{p}_\ell \neq \tilde{p}_{\ell'}$. Such an approach is appealing as it naturally follows from the model construction.

Although this testing procedure is theoretically well-justified, there might be few practical difficulties that are worth emphasizing. Consider the example in Scenario II of Section 5.1 in Camerlenghi *et al.*, in which there are two mixtures of two normal distributions with a common component. The two distributions can be made equal either allowing the weight of the idiosyncratic component to be zero, or having arbitrary weights and letting the distribution-specific components to have the same parameters. The former case can easily be encountered. In fact, when the parameter γ is large enough one has that $w_\ell \approx 0$, in turn implying that $\tilde{p}_\ell \approx \tilde{p}_S$. This statement is formalized in the following lemma, whose proof is omitted.

Lemma 1. *Let $(\tilde{p}_1, \dots, \tilde{p}_d)$ be a latent nested Dirichlet process of (1)–(2). Then $\tilde{p}_1 = \cdots = \tilde{p}_d$ almost surely, as $\gamma \to \infty$.*

Lemma 1 holds for general LNPs and it has relevant consequences. Strictly speaking, it implies that homogeneity among populations is recovered as limiting case when $\gamma \to \infty$, regardless of the ties occurring in the Pólya-sequence of (1). Besides, (2) suggests that homogeneity between two groups (i.e. $\tilde{p}_\ell = \tilde{p}_{\ell'}$) is attained exactly whenever $p_\ell = p_{\ell'}$ but also approximately if $w_\ell \approx 0$. This could affect the rationale underlying the testing procedure, because an LNP model may struggle in discriminating between the case of two identical latent distributions ($p_\ell = p_{\ell'}$) and that of two similar, yet different, random probability measures ($w_\ell \approx 0$). Note that this issue is specific to the LNP, since nested processes correspond to the case $w_\ell = 1$.

As a consequence, if an LNP is employed for testing purposes, the probability of homogeneity $\mathbb{P}(\tilde{p}_\ell = \tilde{p}_{\ell'} \mid \boldsymbol{X})$ might be deflated, possibly leading to biased decisions. Hence, we recommend to select the parameter γ with great care. In contrast, if the LNP were used for density estimation, these considerations would not be a concern.

References

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 1(2): 353–355. MR0362614. 1346

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of normalized random measures with independent increments." *The Annals of Statistics*, 31(2): 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 1346

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association*, 103(483): 1131–1144. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 1346