Bayesian Analysis (2025) **TBA**, Number TBA, pp. 1

Invited Discussion

Beatrice Franzolini* and Giovanni Rebaudo†

We congratulate the authors on their deep theoretical and methodological contributions, which provide a novel understanding as well as new tools for the challenging problem of Bayesian nonparametric regression and covariate-dependent clustering.

Arguably, a substantial portion of methodological advancements within the Bayesian nonparametric literature, since the seminal paper of Ferguson (1973), can be framed as extensions or generalizations of alternative characterizations of the Dirichlet process. Among these characterizations, the stick-breaking (SB) representation (Sethuraman, 1994; Ishwaran and James, 2001) stands out as particularly influential, having been extended to incorporate diverse distributions for the *splitting variables* resulting in different or more general classes of processes, e.g., Rodriguez and Dunson (2011); Favaro et al. (2012); Rigon and Durante (2021). The authors propose a compelling avenue for a different generalization of the standard SB process by highlighting that, beyond specifying distributions for the splitting variables, another implicit yet crucial assumption that can be relaxed is the *lopsided* structural arrangement of the associated representation tree. Such a perspective, on one hand, provides a different take and understanding of existing SB processes, and on the other, lays out novel tractable strategies to define discrete random probability measures. The latter appears particularly useful when the problem at hand requires the construction of an entire collection of random probability measures $(G_x, x \in \mathcal{X})$.

The authors define a generalized class of SB processes (obtained by an arbitrary dyadic tree structure and imposing a tail-free condition) and study the corresponding class of dependent discrete nonparametric priors with common atoms and covariate-dependent SB weights. More precisely, the authors consider the family of discrete distributions $G_x = \sum_{k=1}^{K} W_{x,k}\delta_{\theta_k}$, where $\theta_k \overset{iid}{\sim} G_0$ independent of the weights, and the weights follow a general dyadic-tree SB structure under a tail-free condition. The class provides a theoretical framework for studying and understanding existing covariate-dependent SB priors from a novel, interesting level of generality.

Importantly, in Theorem 1, the authors derive closed-form expressions for the joint moments that are crucial for prior elicitation in terms of marginal and bivariate dependence assumptions. In the next Section, we highlight the connection of these results to the recent theory of multivariate species sampling processes (mSSP) (Franzolini et al., 2025), which provides additional insights and results for studying more general dependence in the authors' covariate-dependent mixtures. The remaining sections of the present discussion contain some brief comments on interesting possible further investigations about the tree-SB framework.

*Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy beatrice.franzolini@unibocconi.it

†University of Torino and Collegio Carlo Alberto, Torino, Italy giovanni.rebaudo@unito.it

1 Connection with theory of mSSP

The results presented in Theorem 1 can be connected to the theory of mSSP introduced by Franzolini et al. (2025). Consider an arbitrary finite-dimensional projection $(G_x : x \in \mathcal{X}_0)$ of $(G_x : x \in \mathcal{X})$, where $\mathcal{X}_0 := \{x_1, \ldots, x_J\} \subseteq \mathcal{X}$ is a finite subset of covariate values. Such finite-dimensional projections $(G_x : x \in \mathcal{X}_0)$ constitute instances of mSSP. Specifically, this includes the marginal univariate and bivariate projections considered in Theorem 1, thereby linking the result in Theorem 1 for tree-SB processes to the mSSP framework. This connection provides a further understanding and validation of the results for tree-SB processes.

The class of mSSP fully characterizes partially exchangeable partitions induced by equivalence relations (ties) within general partially exchangeable arrays. The distribution of these partitions is described by the partially exchangeable partition probability function (pEPPF), which plays a central role in analyzing dependent discrete random processes, including finite-dimensional projections of tree-SB priors.

More explicitly, consider an infinite-dimensional partially exchangeable array $(Y_{i,x} : x \in \mathcal{X}_0, i \in \mathbb{N})$, with groups identified by covariate values $x \in \mathcal{X}_0$, and whose de Finetti measure $\mathcal{L}$ corresponds to the dependent tree-SB law of $(G_x : x \in \mathcal{X}_0)$. Formally, we have:

$$Y_{x,i} \mid (G_x : x \in \mathcal{X}_0) \overset{ind}{\sim} G_x \quad \text{for all } x \in \mathcal{X}_0, i \in \mathbb{N}, \quad (G_x : x \in \mathcal{X}_0) \sim \mathcal{L}.$$

Following Franzolini et al. (2025), the results in Theorem 1 can be expressed in terms of observable quantities, particularly tie probabilities between observations:

$$a_{x,x'} = \Pr(Y_{x,i} = Y_{x',i'}), \quad \text{for any } i, i' \in \mathbb{N}, x, x' \in \mathcal{X}.$$

For instance, if the within-group tie probabilities are equal for groups x and x', namely $\Pr(Y_{x,i} = Y_{x,i'}) = \Pr(Y_{x',j} = Y_{x',j'})$, then equation (9) in Theorem 1 can be rewritten succinctly as

$$\operatorname{corr}\{G_x(A), G_{x'}(A)\} = \frac{\operatorname{corr}(Y_{x,1}, Y_{x',2})}{\operatorname{corr}(Y_{x,1}, Y_{x,2})} = \frac{\Pr(Y_{x,1} = Y_{x',1})}{\Pr(Y_{x,1} = Y_{x,2})} = \frac{\Pr(\text{tie across groups})}{\Pr(\text{tie within a group})}.$$

This formulation provides clarity by translating the theoretical results obtained in terms of the random probabilities representation into simple probabilistic statements regarding observable quantities derived via marginalization of the pEPPF.

Indeed, given an arbitrary vector of dependent random measures $(G_x : x \in \mathcal{X}_0)$, not necessarily tree-SB nor mSSP, we can define the pEPPF as

$$\text{pEPPF}_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) = \mathbb{E}\left[\int_{\mathbb{Y}_*^D} \prod_{d=1}^{D} G_{x_1}(\mathrm{d}x_d)^{n_{1,d}} \ldots G_{x_J}(\mathrm{d}x_d)^{n_{J,d}} \right],$$

under the constraint that $\sum_{d=1}^{D} n_{j,d} = I_j$, with I_j sample size observed from G_{x_j} for each $j = 1, \ldots, J$, $n = \sum_{j=1}^{J} I_j$, and where $\mathbb{Y}$ denotes the space in which the $Y_{x_j,i}$'s

take values, while $\mathbb{Y}_*^D$ denotes the subset of $\mathbb{Y}^D$ consisting of vectors with all distinct entries. The pEPPF characterizes arbitrary partitions induced by partially exchangeable arrays and can be characterized by the class of mSSP (Franzolini et al., 2025). Note that $\text{pEPPF}_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J)$ returns the probability of a given partition, as a function of the sufficient statistics, that is the vector of frequency counts $\boldsymbol{n}_j = (n_{j,1}, \ldots, n_{j,D})$, where $n_{j,d}$ indicates the number of observations in the jth group that coincide with the dth distinct value, indexed according to the order of arrival by groups.

We can now rewrite (9) in Theorem 1 as a marginalization of the pEPPF when available.

$$\text{corr}(G_x(A), G_{x'}(A)) = \frac{\text{pEPPF}_1^{(2)}(1,1)}{\sqrt{\text{EPPF}_{x,1}^{(2)}(1)}\sqrt{\text{EPPF}_{x',1}^{(2)}(1)}},$$

where the EPPF is the exchangeable partition probability function that is a special case of the pEPPF with a single covariate level (i.e., under the exchangeability).

Moreover, under the a.s. discrete mSSP, such as the dependent tree-SB, the pEPPF can be expressed as a simpler function of the weights:

$$\text{pEPPF}_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) = \mathbb{E}\left[\sum_{h_1 \neq \ldots \neq h_D} \prod_{j=1}^{J} \prod_{d=1}^{D} W_{x_j,h_d}^{n_{j,d}}\right].$$

obtaining, indeed, the definition of $a_{x,x'}$ and $a_{x,x}$ by the authors when considering $\text{pEPPF}_1^{(2)}(1,1)$ and $\text{EPPF}_{x,1}^{(2)}(1)$.

General novel results concerning dependence in the framework of mSSP can further provide valuable insights into addressing a key issue highlighted by the authors in their discussion:

> *[. . .] future work involves quantifying the prior dependence between random measures beyond the pairwise linear correlation [. . .].*

For instance, as demonstrated in Franzolini et al. (2025), the $\text{corr}\{G_x(A), G_{x'}(A)\}$ constitutes an effective index of dependence, and thus of the borrowing of information for a large class of dependent process that include the dependent tree-SB considered by the authors. Indeed, not only is it a global index in $[0,1]$ not depending on the specific set A considered, but

- $\text{corr}(G_x(A), G_{x'}(A)) = 1$ if and only if $G_x = G_{x'}$ a.s., i.e. full exchangeability;
- $\text{corr}(G_x(A), G_{x'}(A)) = 0$ if and only if G_x and $G_{x'}$ are independent.

in such a class. Consequently, within this class, the correlation is always reflective of the underlying dependence structure. Due to the intrinsic properties of these processes, the correlation measure is not subject to the typical limitations associated with linear measures of dependence, and, in particular, a correlation of zero cannot arise unless there is a complete absence of dependence.

Moreover, general dependence beyond pairwise dependence can be expressed in terms of higher-order moments of dependent tree-SB processes evaluated on (measurable) sets. These quantities can be meaningfully interpreted in terms of simple observable quantities, that is, the random number of observed species within and across groups, similarly to what has been shown above for the correlation. We refer to Franzolini et al. (2025) for further details.

2 Choice of K

As the authors state in the discussion:

> *One limitation of the balanced tree model investigated in this work is that we assumed the tree is truncated at a fixed maximum depth.*

The choice of K is crucial from both a modeling and theoretical perspective. For instance, roughly speaking, increasing the truncation level K increases the probability of observing both more clusters and with smaller frequency for each cluster (contrary to the lopsided counterpart). Moreover, in the nonparametric case it is necessary to take $K = \infty$ (without degenerating to a diffuse fixed G_0) or to allow K to be a random, unbounded integer in order to achieve full support for the random probability measure. In addition, allowing K to be infinite (or random and unbounded) is also desirable for theoretical guarantees. In particular, it opens the possibility of achieving frequentist consistency of the Bayesian mixture to a *true* data-generating density (see, e.g., Ghosal and van der Vaart, 2017) or a *true* number of components (see, e.g., Nobile, 1994; Miller and Harrison, 2018; Ascolani et al., 2023; Miller, 2023; Zeng et al., 2023), that might be investigated in future work.

3 Alternative tree structures for existing processes

The significance of the tree-based perspective introduced by the authors is twofold. First, the authors provide a general framework for constructing novel classes of random probability measures, which can effectively address notable limitations such as high correlation and heightened posterior uncertainty inherent in default choices of lopsided stick-breaking (SB) processes, while still maintaining analytical and computational tractability. Second, the tree framework provides insights into known processes defined via SB. This second perspective possibly also offers a promising pathway to derive new convenient representations of existing processes. Although the present study extensively explores the first direction, we anticipate that future research leveraging this framework may also yield alternative representations, potentially more analytically transparent or computationally advantageous, for existing processes that currently lack analytical or computationally tractable representations.

To move in this direction, consider for simplicity a single process $G \overset{a.s.}{=} \sum_{k=1}^{K} W_k \delta_{\theta_k}$, with a certain law. Suppose we aim to determine the distribution of the splitting variables for this process, given a particular tree structure τ. A few natural questions arise. Given a certain tree structure τ with a sufficient number of terminal leaves, is it always

possible to define the distribution of the splitting variables that produces the given distribution for the weights $(W_k)_{k\geq 1}$ if one removes the tail-free condition? If such a representation exists, can the tail-free condition for a generic tree structure τ be linked to some (specific) theoretical property of the processes that satisfy it?

Although we acknowledge that the breadth of these questions makes them challenging to answer, we believe that further research aimed at characterizing the class of processes available for a given tree structure τ could yield important additional insights into the proposed perspective and, possibly, novel representations of processes that currently lack analytically or computationally tractable representations.

4 Conclusion

The authors' contribution represents a significant step forward in the development of covariate-dependent Bayesian nonparametric priors. By formalizing a general class of stick-breaking models with common atoms and tractable dependence structures, they offer both a unifying theoretical lens and new modeling tools for regression and clustering tasks. We look forward to seeing further theoretical developments and applied investigations building on this rich framework.

References

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2023). "Clustering consistency with Dirichlet process mixtures." *Biometrika*, 110: 551–558. 4

Favaro, S., Lijoi, A., and Prünster, I. (2012). "On the stick-breaking representation of normalized inverse Gaussian priors." *Biometrika*, 99: 663–674. 1

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Ann. Stat.*, 1: 209–230. 1

Franzolini, B., Lijoi, A., Prünster, I., and Rebaudo, G. (2025). "Multivariate species sampling models." *Preprint at arXiv: 2503.24004*. 1, 2, 3, 4

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Univ. Press. 4

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *J. Am. Stat. Assoc.*, 96: 161–173. 1

Miller, J. W. (2023). "Consistency of mixture models with a prior on the number of components." *Depend. Model.*, 11: 1–9. 4

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components." *J. Am. Stat. Assoc.*, 113: 340–356. 4

Nobile, A. (1994). "Bayesian Analysis of Finite Mixture Distributions." Ph.D. thesis, Carnegie Mellon Univ. 4

Rigon, T. and Durante, D. (2021). "Tractable Bayesian density regression via logit stick-breaking priors." *J. Stat. Plan. Inference*, 211: 131–142. 1

Rodriguez, A. and Dunson, D. B. (2011). "Nonparametric Bayesian models through probit stick-breaking processes." *Bayesian Anal.*, 6: 10–1214. 1

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Stat. Sin.*, 4: 639–650. 1

Zeng, C., Miller, J. W., and Duan, L. L. (2023). "Consistent model-based clustering: using the quasi-Bernoulli stick-breaking process." *J. Mach. Learn. Res.*, 24: 1–32. 4