Spatial neutral-to-the-left stochastic processes

Vittorio Romaniello

March 29, 2020

1 Summary

A sequence of random variables $X = (X_1, \ldots, X_k) \in \mathbb{R}_+^k$ is said to be neutral-to-the-left (NTL) if the increments

$$R_j = \frac{X_j}{\sum_{i=1}^{j} X_i}, \quad j = 1, \ldots, k$$

form a sequence of mutually independent random variables in $[0, 1]$ [Blo+18]. Such processes are common in statistics and can be used to describe, for instance, the generating process of left censored data or the asymptotic degree distribution in preferential attachment random graphs [BO17]. In the statistics and probability literature and, especially, in Bayesian nonparametrics, NTL processes have not received as much attention as their analogous counterpart, neutral-to-the-right (NTR) processes [1]. The disparity in the development of theory for NTL and NTR processes can be attributed to two factors:

1. NTR processes are more common than NTL processes, therefore more focus has been put on the development of theory for the former.

2. Since NTL and NTR are symmetric opposites [Blo+18], an NTL process can be transformed into an NTR for the analysis and back to an NTL after the analysis.

However, the increased attention received by the study of random graphs as well as the interest to model streaming data has made NTL processes more popular and highlighted the need to develop a theory for NTL processes. In fact, the approach used so far in the study of NTL processes is cannot be applied to situations where data is collected in a streaming fashion. Transforming and NTL into an NTR process requires knowledge of the full dataset hence, analysis of NTL processes of streaming data cannot be performed with the current approach.

To overcome this issue and expand existing literature on NTL processes, we present here the spatial neutral-to-the-left (SPNTL) stochastic process. SPNTL enables direct Bayesian modeling of NTL data and can be used as prior for Bayesian nonparametric models. In this paper, following the theoretical analysis of [Jam06] for spatial neutral-to-the-right (SPNTR) processes, we derive a characterization of the posterior distribution of SPNTL using results from Poisson partition calculus [Jam05].

While the results in our paper are important to develop a theory for NTL processes, our work does not cover all theoretical contributions of [Jam06] for the setting of NTL processes. Deriving additional results for SPNTL is left as future work and discussed more in detail in Section 2 together with possible applications.

[1] Reversing an NTL process gives an NTR process.

2 Mini-proposals

2.1 Proposal 1: NTL species sampling and generalized Chinese restaurant process

The work of [Jam06], that guided our analysis, was motivated by the observation that despite the numerous applications suitable for NTR priors, NTR processes were not being used as extensively as they could have been. The Dirichlet process constituted the only exception and it is still ubiquitously used in Bayesian nonparametrics applications.

[Jam06] identified several reasons for the absence of NTR priors from the Bayesian nonparametrics literature and addressed them in his work.

Firstly, he noticed that the NTR process of [Dok74] was defined only on the real line, $\mathbb{R}$, therefore restricting its application to processes in the real space. Hence, NTR processes could not be applied to situations where inference on more complex spaces was required. To address this limitation, [Jam06] proposed the spatial neutral-to-the-right process (SPNTR), an extension of NTR processes to arbitrary Polish spaces. From SPNTR, [Jam06] derived a new class of random probability measures called NTR species sampling models. This work focuses on characterizing SPNTL processes similarly to [Jam06], however, the possibility to derive analogous NTL species sampling models is not explored and could be further investigated following the analysis of [Jam06].

Secondly, and proba.bly the most important problem addressed in [Jam06], is that NTR processes are, in general, not tractable and there is no simple way to sample from them, limiting their practical application. On the other hand, Dirichlet processes can easily be sampled through the Chinese restaurant process. [Jam06] derives a generalized Chinese restaurant process sampling scheme for SPNTR processes, making their implementation more approachable. Our work has not developed a similar sampling scheme for SPNTL.

We believe that exploring the two research directions presented in this proposal is essential to make SPNTL processes available for practical use. The analysis could be performed following the work of [Jam06] and can bring forward the theoretical development of NTL processes.

2.2 Proposal 2: Application to streaming data

As mentioned in Section 1, the approach of transforming an NTL into an NTR process, performing the analysis as if it were an NTR process and transforming the NTR back to NTL requires knowledge of the full dataset and cannot be applied to streaming data. SPNTL allow to overcome this issue providing a way to model data directly as a NTL process. In this paper we focused on deriving the theoretical results for SPNTL, however, it is interesting to observe the advantage obtained by directly modeling the NTL process.

Furthermore, it would be interesting to use SPNTL to model streaming data and compare inference performance to established methods. In particular, we think that by being able to model not only the temporal component of the process but also the action on more complex spaces SPNTL could lead to improved performance.

Testing performance of SPNTL models could highlight practical weaknesses of the models and foster ideas for new research directions and theoretical developments.

Data that could be used for this project could come from the recommender systems literature, where they could perhaps be used to address the cold start problem, a well-known problem for recommender systems. Another application could be in the context of random graphs [BO17; Blo+18].

3 Project report

Lastly, to address the issues above, [Jam06] used results from Poisson partition calculus [Jam05]. Although the results from Poisson calculus come from a previous work [Jam05], the approach greatly simplified proofs and allows for a relatively simple generalization of NTR processes. Approaches used in literature prior to this work derived sampling distributions for Dirichlet processes using combinatorial arguments [Ant74; Pit+02]. Adopting a similar strategy in generalizing NTR processes would result challenging and would likely not be viable.

goes beyond deriving a posterior characterization for the SPNTR. In fact, James' work was motivated by a different observation and

Since numerous contributions were made by [Jam06]

Neutral-to-the-right (NTR) stochastic processes [Dok74] were introduced as priors for Bayesian nonparametrics models of right censored data (e.g survival analysis). In Bayesian statistics, using priors specifically designed for the problem at hand is renowned to be beneficial for effective inference. Hence, having the possibility to improve the descriptiveness of models for right censored data using NTR priors should be seen as a great advantage.

References

[Ant74] C. E. Antoniak. "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". In: *The annals of statistics* (1974), pp. 1152–1174.

[BO17] B. Bloem-Reddy and P. Orbanz. "Preferential attachment and vertex arrival times". In: *arXiv preprint arXiv:1710.02159* (2017).

[Blo+18] B. Bloem-Reddy et al. "Sampling and inference for Beta Neutral-to-the-Left models of sparse networks". In: *arXiv preprint arXiv:1807.03113* (2018).

[Dok74] K. Doksum. "Tailfree and neutral random probabilities and their posterior distributions". In: *The Annals of Probability* (1974), pp. 183–201.

[Jam05] L. F. James. "Poisson process partition calculus with an application to Bayesian Lévy moving averages". In: (2005).

[Jam06] L. F. James. "Poisson calculus for spatial neutral to the right processes". In: *The annals of Statistics* 34.1 (2006), pp. 416–440.

[Pit+02] J. Pitman et al. *Combinatorial stochastic processes*. Tech. rep. Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for . . ., 2002.