Spatial neutral-to-the-left stochastic processes

Vittorio Romaniello

March 30, 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

In this section we follow the analysis of [Jam06] to derive a posterior distribution for the SPNTL process. In Section 3.1 we give a rigorous definition of a NTL process and derive a formula of the random distribution function in terms of a stochastic process, similarly to Thm. 3.1 of [Dok74]. Section 3.2 introduces the cumulative birth measure, the random measure we will use to specify a SPNTL process. Section 3.3 outlines the SPNTL process and in Section 3.4 we detail the posterior analysis and derive the main result of the paper using results from Poisson partition calculus [Jam05; Jam06].

3.1 Neutral to the left stochastic process

The Bayesian analysis of SPNTL is based on NTL stochastic processes. [Dok74] provides a framework to analyse NTR processes and characterize the random distribution function in terms of a Levy process. The NTR characterization, as well as the relation between the random distribution function and a Levy process is essential to obtain the results in [Jam06]. Therefore, we give here a similar characterization for NTL processes as this will aid the posterior analysis of SPNTL.

Given a sequence of random variables $X = (X_1, \ldots, X_k)$, we say that X is NTL if the increments form a sequence, R, of mutually independent random variables in $[0, 1]$ [Blo+18], that is

$$R = \left(1, \frac{X_2}{X_1 + X_2}, \ldots, \frac{X_j}{\sum_{i=1}^{j} X_i}, \ldots, \frac{X_k}{\sum_{i=1}^{k} X_i}\right)$$

is a vector of mutually independent random variables on $[0, 1]$. Let $V_1, \ldots, V_k$ be nonnegative independent random variables in $[0, 1]$ then R can be defined as

$$R = (1, V_2, \ldots, V_k)$$

and

$$\left(1, \frac{X_2}{X_1 + X_2}, \ldots, \frac{X_j}{\sum_{i=1}^{j} X_i}, \ldots, \frac{X_k}{\sum_{i=1}^{k} X_i}\right) =_d (1, V_2, \ldots, V_j, \ldots, V_k) \tag{1}$$

where $=_d$ indicates equality in distribution a.s. From Equation (1) it is possible to derive the distribution of X in terms of $V = (V_1, \ldots, V_k)$ as an "inverse" stick-breaking process like in [BO17].

$$(X_1, X_2, \ldots, X_j, \ldots X_k) =_d \left(\prod_{i=2}^{k}(1 - V_i), V_2 \prod_{i=3}^{k}(1 - V_i), \ldots, V_j \prod_{i=j+1}^{k}(1 - V_i), \ldots, V_k\right)$$

where we set $V_1 = 1$.

The increments in Equation (1) can be written equivalently in terms of a random distribution function F corresponding to a random probability P such that $F(t) = P([-\infty, t]) \in [0, 1]$ for $t \in \mathbb{R}_+$. The random distribution function is a stochastic process that satisfies

1. F is a.s. non-decreasing
2. $\lim_{t \to -\infty} F(t) = 0$ a.s., $\lim_{t \to \infty} F(t) = 1$ a.s, and
3. $\lim_{s \to t^+} F(s) = F(t)$ for each $t \in \mathbb{R}_+$

as stated in [Dok74].

The increments in Equation (1) can be rewritten in terms of the stochastic process F as

$$\left(\frac{F(t_1)}{F(t_1)}, \frac{F(t_2) - F(t_1)}{F(t_2)}, \ldots, \frac{F(t_k) - F(t_{k-1})}{F(t_k)}\right) =_d (1, V_2, \ldots, V_k) \tag{2}$$

for $0 < t_1 < t_2 < \ldots < t_k < \infty$. The random distribution F is essentially a random cumulative density function (CDF) and its stochastic process corresponds to a special type of Levy process called *subordinator* [Orb12].

Levy process [Pap08] A stochastic process $Z = \{Z(t) : t \in \mathbb{R}_+\}$ is said to be a Levy process if it satisfies the following properties:

1. $Z(0) = 0$
2. For any $0 < t_1 < t_2 < \ldots < t_k < \infty$, $Z(t_2) - Z(t_1), \ldots, Z(t_k) - Z(t_{k-1})$ are independent.
3. For any $s < t$, $Z(t) - Z(s) =_d Z(t - s)$
4. For any $\epsilon > 0$ and $t \in \mathbb{R}_+$, $\lim_{s \to t} P(|Z(t) - Z(s) > \epsilon|) = 0$

Given the correspondence between the random CDF and a Levy process, we can find a function of a Levy process, Z, such that the random distribution F satisfies the NTL properties. The need to express F as a function of Z is motivated by the fact that working with a Levy process simplifies some derivations. In Theorem 1 we define F as a function of a Levy process for NTL process, the analogous of Theorem 1 for NTR processes can be found in [Dok74].

Theorem 1 $F_{t_k}(t)$ is a random distribution function neutral-to-the-left if and only if it has the same probability distribution as

$$\exp(Z(t) - Z(t_k)), \quad t \in (0, t_k] \tag{3}$$

for some Levy process Z.

Proof: We start by proving that by writing F_{t_k} as in Equation (3) we obtain independent increments. For any increment $I[t_{j-1}, t_j]$ of a NTL process,

$$\frac{F_{t_k}(t_j) - F_{t_k}(t_{j-1})}{F(t_j)} = 1 - \frac{F_{t_k}(t_{j-1})}{F_{t_k}(t_j)} = 1 - \exp\left(-(Z(t_j) - Z(t_{j-1}))\right) \tag{4}$$

since increments of a Levy process are independent, it follows that increments of a NTL generated by Equation (3) are independent.

We now show that the process, $Z(t)$, following from the definition of F has independent increments, hence proving it is a Levy process. We first write a general expression for $F_{t_k}(t_j)$ derived from Equation (2). First note that

$$\frac{F_{t_k}(t_k) - F_{t_k}(t_{k-1})}{F_{t_k}(t_k)} = V_k \implies F_{t_k}(t_{k-1}) = (1 - V_k)$$

recursively we obtain

$$F_{t_k}(t_j) = \prod_{i=j+1}^{k} (1 - V_i) \tag{5}$$

The increments of the stochastic process $Z(t)$ can be written as

$$\begin{aligned} Z(t_j) - Z(t_{j-1}) &= \log(F_{t_k}(t_j)) - \log(F_{t_k}(t_{j-1})) = \log\left(\frac{F_{t_k}(t_j)}{F_{t_k}(t_{j-1})}\right) \\ &= \log\left(\frac{\prod_{i=2}^{j-1}(1 - V_i)}{\prod_{i=2}^{j}(1 - V_i)}\right) = -\log(1 - V_j) \end{aligned}$$

hence showing that the increment at time t_j only depends on a function of V_j and since V_i for $i = 1, \ldots, k$ are independent, the increments of $Z(t)$ are independent. This shows that $Z(t)$ is a Levy process. □

Note that it is important to define F on a specific interval to ensure it represents a proper CDF. However, a fixed time interval is by no mean restrictive and the results derived in the following hold for any time interval. In fact, the random CDF in Equation (3) can be normalized to any interval as

$$F_{t_k}(t) = \frac{F_{t_{k-1}}}{\exp(Z(t_k) - Z(t_{k-1}))} \quad t \in (0, t_k]$$

The definition of F represents a stochastic process on $\mathbb{R}_+ \times \mathcal{B}(\mathbb{R}_+)$ and is valid for any time interval. Hence, to simplify notation in the rest of the analysis we omit the subscript of F unless necessary.

3.2 Cumulative birth measure

To analyse NTR processes [Jam06] uses the approach of [Hjo90] and works with the cumulative hazard measure Λ, a completely random measure that directly relates to the NTR increments. For NTL processes we define a similar object and name it cumulative birth measure

$$B(ds) = F_{t_k}(ds)/F_{t_k}(s+), \quad s \in (0, t_k] \tag{6}$$

where $ds = (s+) - (s-)$ and $F_{t_k}(s+) = P(T \leq s+)$ with T a random variable following the distribution F. B is a completely random measure. Interestingly, defining the cumulative birth measure as in Equation (6) lets us relate random jumps in the Levy process Z to random jumps in B in exactly the same way as done in [Jam06] for Z and Λ.

Proposition 1 A random jump J_j in B taking values in $[0, 1]$ corresponds to a random jump $-\log(1 - J_j)$ in Z taking values in $\mathbb{R}_+$.

Proof: The proof is simple and follows equating B and J_j and using Equation (4). Let t_j and t_{j-1} denote the interval of the jump J_j and denote $dj = t_j - t_{j-1}$ then

$$B(dj) = \frac{F(dj)}{F(t_j)} = 1 - \exp(-(Z(t_j) - Z(t_{j-1}))) = J_j \implies Z(t_j) - Z(t_{j-1}) = -\log(1 - J_j) \quad \square$$

Using the cumulative birth measure we can write

$$P(T \in ds|F) = F(ds) = B(ds)F(s+) = B(ds)\exp(Z(s+) - Z(t_k))$$

3.3 Spatial neutral to the left process

The previous sections provided the background necessary to define the SPNTL process. Since our work extends [Jam06], we borrowed James' approach and notation in several parts.

The SPNTL process is motivated by the observation that the NTL process does not allow inference on spaces more complex than the real line however, complex spaces often appear in practice. In order to work with more complex spaces, we extend the space on which the random distribution F is defined to an arbitrary Polish space $\mathcal{S} = \mathbb{R}_+ \times \mathcal{B}(\mathbb{R}_+) \times \mathcal{X}$ with $\mathcal{X}$ an arbitrary Polish space. We denote by (R, T, X) the elements of the Polish space $\mathcal{S}$ that have distribution $F_r(ds, dx)$ for $(r, s, x) \in \mathcal{S}$. The extension to the Polish space $\mathcal{S}$ is such that $F_r(ds, \mathcal{X})$ is a NTL process. To avoid carrying over the subscript in F, in the rest of the analysis we assume r fixed and deal only with $(s, x) \in \mathcal{S}_{-r}$.

In his analysis of SPNTR [Jam06] used results from Poisson partition calculus. This, significantly simplified proofs as combinatorial arguments, a common but challenging approach for similar proofs (see e.g. [Ant74; Pit+02]), could be avoided. To apply Poisson calculus for SPNTL processes, we extend Z and B to completely random measures on $\mathcal{S}_{-r}$ using a representation in terms of a Poisson random measure, similarly to [Jam06]. Let N denote a Poisson random measure on the Polish space $\mathcal{W} = [0, 1] \times \mathcal{S}_{-r}$ with mean intensity

$$E[N(du, ds, dx)|\nu] = \nu(du, ds, dx) = \rho(du|s)B_0(ds, dx)$$

where ρ is a Levy density that determines the conditional distribution of the jumps of B and Z satisfying $\int_0^1 u\rho(du) = 1$ as in [Jam06] and $B_0(ds, dx) = F_0(ds, dx)/F_0(s+)$ is the birth measure on $\mathcal{S}_{-r}$ with F_0 denoting a prior on F in $\mathcal{S}_{-r}$ and $F_0(s+) = F_0(s+, \mathcal{X})$.

Moreover, we denote the law of N with intensity ν as $P(dN|\nu)$ and the Laplace functional of N as

$$E\left[e^{-N(f)}|\nu\right] = \int_{\mathbb{M}} e^{-N(f)} P(dN|\nu) = e^{-\mathcal{G}(f)}$$

where for any positive f, $N(f) = \int_{\mathcal{W}} f(x)N(dx)$ and $\mathcal{G}(f) = \int_{\mathcal{W}}(1 - e^{-f(x)})\nu(dx)$ and $\mathbb{M}$ is the space of boundedly finite measures on $\mathcal{W}$. The definitions above are the same as those in [Jam06], however, we presented them here since they will be necessary for the rest of the analysis.

From the specifications above we have that $B(ds, dx) = \int_0^1 uN(du, ds, dx)$ is a completely random measure on $\mathcal{S}_{-r}$ with $E[B(ds, dx)] = B_0(ds, dx)$ and the corresponding for the Levy process Z, $Z(ds, dx) = \int_0^1 [-\log(1-u)]N(du, ds, dx)$. Furthermore, we define $Z(t+) = \int_{\mathcal{W}} [-\mathbb{1}(s < t)\log(1-u)]N(du, ds, dx)$. Using these definitions we can specify a SPNTL random probability measure on $\mathcal{S}_{-r}$ as

$$P(T \in dt, X \in dx|F) = F(dt, dx) = F(t+)B(dt, dx) \tag{7}$$

and $E[F(dt, dx)] = F_0(dt, dx)$.

3.4 Posterior analysis

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.

[Hjo90] N. L. Hjort. "Nonparametric Bayes estimators based on beta processes in models for life history data". In: *the Annals of Statistics* 18.3 (1990), pp. 1259–1294.

[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.

[Orb12] P. Orbanz. "Lecture notes on bayesian nonparametrics". In: *Journal of Mathematical Psychology* 56 (2012), pp. 1–12.

[Pap08] A. Papapantoleon. "An introduction to L\'{e} vy processes with applications in finance". In: *arXiv preprint arXiv:0804.0482* (2008).

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