Pólya urn models and neutral theory of biodiversity

February 18, 2025

Supervision

- Jean Peyhardi, Institut Montpellierain Alexander Grothendieck, Université de Montpellier <jean.peyhardi@umontpellier.fr>
- Stéphane Robin, Laboratoire de Probabilité, Statistiques et Modélisation, Sorbonne Université <stephane.robin@sorbonne-universite.fr>

1 PhD Context

Neutral theory in ecology

The unified neutral theory of biodiversity introduced by Hubbell et al. (2001) emphasizes the importance of stochastic processes in ecological community structure, and has challenged the traditional niche-based view of ecology. It has challenged classical theories of species diversity by showing that patterns of species diversity similar to those observed in nature can be obtained from an extremely simplified model of community dynamics where each dying individual is immediately replaced by a new individual (zero-sum) and all individuals of all species are ecologically identical (neutrality).

Formally, the focus is on the stationary distribution of a multivariate birth and death process $\boldsymbol{N}(t) = (N_1(t), \dots, N_J(t))$, where $N_j(t)$ is the abundance of species j at time , and J is the total number of species. Under the zero-sum assumption (i.e., fixed sum $|\boldsymbol{N}(t)| = n$) - and mild hypothesis on birth and death rates - Hubbell et al. (2001) showed that the multivariate distribution of $\boldsymbol{N}$ given $|\boldsymbol{N}| = n$ (at equilibrium) is a Dirichlet multinomial distribution $\mathcal{DM}_n(\theta_1, \dots, \theta_J)$. Haegeman and Etienne (2008) proposed to relax the zero-sum assumption in order to obtain a more realistic model and they also find a Dirichlet multinomial distribution at equilibrium for the conditional distribution of $\boldsymbol{N}$ given $|\boldsymbol{N}| = n$. But they added the assumption of independence between species to obtain an analytic formula of the sum distribution. Indeed, in this case, it is sufficient to study the abundance distribution for each species and then use the convolution. They find a negative binomial for each species and then use the closure under convolution to obtain also a negative binomial distribution for the sum. In fact, Haegeman and Etienne (2017) proposed to replace the zero-sum assumption by the assumption of independence between species abundances. In a recent work, Peyhardi et al. (2024) proposed to generalize this approach in two ways. Firstly, they relaxed both, the zero-sum assumption and the independence assumption. Secondly, they consider the enlarged family of Pólya splitting distributions, that includes the Dirichlet multinomial with negative binomial sum, as a special case. The Pólya urn models offer a nice framework to deal with joint species distribution models for multi-species abundance data, under the neutral theory of biodiversity.

Mathematical framework

The class of Pólya splitting distributions, introduced by Peyhardi and Fernique (2017) and Jones and Marchand (2019), is defined as compound distributions $\boldsymbol{N} \sim \mathcal{P}_n^{[c]}(\boldsymbol{\theta}) \wedge_n \mathcal{L}$, meaning that the sum $|\boldsymbol{N}|$ follows the univariate distribution $\mathcal{L}$ and $\boldsymbol{N}$ given $|\boldsymbol{N}| = n$ follows the multivariate Pólya distribution. Let us briefly recall the definition of a multivariate Pólya distribution in terms of urn models.

Pólya urn model One urn initially contains θ_j balls of the color j for $j = 1, \ldots, J$. At each draw, one ball is drawn at random and then replaced with c additional balls of the same color, where $c \in \{-1, 0, 1\}$. This procedure is repeated n times and focus is made on the multivariate count $\boldsymbol{N} = (N_1, \ldots, N_J)$ of drawn balls for each color. Knowing the number n of draws, the conditional count distribution of $\boldsymbol{N}$ given $|\boldsymbol{N}| = n$ is known as the multivariate Pólya distribution, denoted by $\mathcal{P}_n^{[c]}(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta_c^J$ (where $\Theta = \mathbb{N}$ for $c = -1$ and $\Theta = \mathbb{R}_+$ otherwise). As expressed by Peyhardi (2023), if we denote

$$a_\theta^{[c]}(n) = \frac{\prod_{k=0}^{n-1}(\theta + ck)}{n!} \mathbb{1}_{\theta + cn \geq 0},$$

then the probability mass function (pmf) of a multivariate Pólya distribution $\mathcal{P}_n^{[c]}(\boldsymbol{\theta})$ takes the following form

$$P_{|\boldsymbol{N}|=n}(\boldsymbol{N} = \boldsymbol{n}) = \frac{1}{a_{|\boldsymbol{\theta}|}^{[c]}(n)} \prod_{j=1}^{J} a_{\theta_j}^{[c]}(n_j).$$

The mulyivariate Pólya distribution turns out to be the multivariate hyergeometric distribution $\mathcal{H}_n(\boldsymbol{\theta})$ when $c = -1$ (without replacement), the multinomial distribution $\mathcal{M}_n(\boldsymbol{\pi})$ when $c = 0$ (with replacement meaning independent draws) and the Dirichlet multinomial distribution $\mathcal{DM}_n(\boldsymbol{\theta})$ when $c = 1$ (with reinforcement).

Remarkable Pólya splitting distributions Properties of Pólya splitting distributions are related to the choice of the sum distribution $\mathcal{L}$. For instance the covariance between N_i and N_j (with $(i, j) \in \{1, \ldots, J\}^2$ and $i \neq j$) is given by

$$\mathrm{Cov}(N_i, N_j) = \frac{\theta_i \theta_j}{|\boldsymbol{\theta}|^2(|\boldsymbol{\theta}| + c)} \left[(\mu_2 - \mu_1^2)|\boldsymbol{\theta}| - c\mu_{(1)}^2 \right], \tag{1}$$

where μ_k is the factorial moment of order k of the sum distribution $\mathcal{L}$. It implies that the sign of covariance between any pair (i, j) is driven by the moments of the sum; see Table 1 for some examples. This table summarizes nine Pólya splitting distributions based on three remarkable choices for the sum distribution $\mathcal{L}$ (Pólya, Power Series and Inverse Pólya); see (Peyhardi, 2023) for details. An important property of these multivariate count distributions is the closure under addition, i.e., for any species pair (i, j) we have

$$\left.\begin{array}{l} N_i \sim \mathcal{L}(\theta_i) \\ N_j \sim \mathcal{L}(\theta_j) \end{array}\right\} \quad N_i + N_j \sim \mathcal{L}(\theta_i + \theta_j).$$

This concept generalizes the closure under convolution since the independence is not required.

Sum \ Split	Hypergeometric $c = -1$	Multinomial $c = 0$	Dirichlet multinomial $c = 1$	Covariance sign
Pólya	$\mathcal{H}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{H}_m(\lvert\boldsymbol{\theta}\rvert, \gamma)$	$\mathcal{M}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{B}_m(p)$	$\mathcal{DM}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \beta\mathcal{B}_m(\lvert\boldsymbol{\theta}\rvert, \gamma)$	negative
Power series	$\mathcal{H}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{B}_{\lvert\boldsymbol{\theta}\rvert}(p)$	$\mathcal{M}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{P}(\lambda)$	$\mathcal{DM}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{NB}(\lvert\boldsymbol{\theta}\rvert, p)$	null
Inverse Pólya	$\mathcal{H}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \beta\mathcal{B}_{\lvert\boldsymbol{\theta}\rvert}(a, b)$	$\mathcal{M}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{NB}(a, p)$	$\mathcal{DM}_n(\boldsymbol{\theta}) \underset{n}{\wedge} \beta\mathcal{NB}(\lvert\boldsymbol{\theta}\rvert, a, b)$	positive

Table 1: Nine remarkable Pólya splitting distributions with different split distributions (columns) and different sum distributions (rows).

Multivariate birth-death processes Peyhardi et al. (2024) exhibited the birth and death rates assumptions that lead to the multivariate Pólya distribution at equilibrium. The master equation

describing the behavior of the multivariate jump process $\boldsymbol{N}(t)$ is given by

$$\frac{\partial p_{\boldsymbol{n}}(t)}{\partial t} = \sum_{j=1}^{J} p_{\boldsymbol{n}-\boldsymbol{e}_j}(t) q_j^-(\boldsymbol{n}-\boldsymbol{e}_j) + p_{\boldsymbol{n}+\boldsymbol{e}_j}(t) q_j^+(\boldsymbol{n}+\boldsymbol{e}_j) - p_{\boldsymbol{n}+\boldsymbol{e}_j}(t)\{q_j^-(\boldsymbol{n}) + q_j^+(\boldsymbol{n})\}$$

where $q_j^-(\boldsymbol{n})$ (resp. $q_j^+(\boldsymbol{n})$) denotes the jumping rate from $\boldsymbol{n}$ to $\boldsymbol{n}-\boldsymbol{e}_j$ (resp. to $\boldsymbol{n}+\boldsymbol{e}_j$), $\boldsymbol{e_j}$ denotes the indicator vector of the j_{th} element and $p_{\boldsymbol{n}}(t) = P\{\boldsymbol{N}(t) = \boldsymbol{n}\}$ (resp. $p_{\boldsymbol{n}} = P(\boldsymbol{N} = \boldsymbol{n})$) denotes the pmf at time t (resp. the pmf at stationary state). Assume that there exists some parameters $c \in \{-1, 0, 1\}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J) \in \Theta_c^J$ and two non-negative functions s^+ and s^- such that

$$\begin{aligned} q_j^+(\boldsymbol{n}) &= s^+(|\boldsymbol{n}|)(\theta_j + cn_j)\mathbb{1}_{\theta_j + cn_j \geq 0}, \\ q_j^-(\boldsymbol{n}) &= s^-(|\boldsymbol{n}|)n_j. \end{aligned} \tag{2}$$

The birth-deah rate $q_j(\boldsymbol{n}) := q_j^+(\boldsymbol{n})/q_j^-(\boldsymbol{n}+\boldsymbol{e_j})$ thus becomes

$$q_j(\boldsymbol{n}) = s(|\boldsymbol{n}|)\frac{\theta_j + cn_j}{n_j + 1}\mathbb{1}_{\theta_j + cn_j \geq 0} \tag{3}$$

where $s(n) = \frac{s^+(n)}{s^-(n+1)}$ for all $n \in \mathbb{N}$. It can be seen that this parametric assumption (3) respects the Kolmogorov's criterion $q_i(\boldsymbol{n})q_j(\boldsymbol{n}+\boldsymbol{e_i}) = q_j(\boldsymbol{n})q_i(\boldsymbol{n}+\boldsymbol{e_j})$, and thus leads to a reversible process. Remarking that $\prod_{k=0}^{n-1}\frac{\theta+ck}{k+1} = a_\theta^{[c]}(n)$ we add the following assumption on $s(n)$ in order to obtain a well defined stationary distribution:

$$\sum_{n\geq 0} a_\theta^{[c]}(n) \prod_{k=0}^{n-1} s(k) < \infty. \tag{4}$$

Theorem 1 *Assume that the hypothesis* (3) *and* (4) *hold then*

- *the stationary distribution of* $\boldsymbol{N}(t)$ *is the Pólya splitting distribution* $\mathcal{P}_n^{[c]}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}$
- $\mathcal{L}$ *is the stationary distribution of a univariate process with birth/death ratio equal to* $q(n) = s(n)r_{|\boldsymbol{\theta}|}^{[c]}(n)$*, more precisely we have* $P(|\boldsymbol{N}| = n) \propto a_{|\boldsymbol{\theta}|}^{[c]}(n)\prod_{k=0}^{n-1} s(k)$.

Using Theorem 1, the parametric form of jumping rates leading to the nine remarkable distributions of Table 1 are easily obtained (see Table 2).

Sum \ Split	Hypergeometric $c=-1$	Multinomial $c=0$	Dirichlet multinomial $c=1$	Covariance sign
Pólya	$\frac{m-\lvert\boldsymbol{n}\rvert}{\gamma-m+\lvert\boldsymbol{n}\rvert+1}$	$\frac{m-\lvert\boldsymbol{n}\rvert}{\gamma}$	$\frac{m-\lvert\boldsymbol{n}\rvert}{\gamma+m-\lvert\boldsymbol{n}\rvert-1}$	negative
Power series	α	α	α	null
Inverse Pólya	$\frac{a+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+b-\lvert\boldsymbol{n}\rvert-1}$	$\frac{a+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+b}$	$\frac{a+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+b+a+\lvert\boldsymbol{n}\rvert+1}$	positive

Table 2: Parametric form of $s(|\boldsymbol{n}|)$ and thus of jumping rates $q_j(\boldsymbol{n})$ leading to the nine remarkable Pólya splitting distribution of Table 1 at equilibrium.

2 PhD subject

Our aim is to describe the boundaries of the neutral theory. Inside the boundaries, the class of model that are closed under addition seems to play a central role. This property could facilitate the addition of speciation in the birth-death process. Outside the boundaries, the class of tree Pólya splitting models, introduced by Valiquette et al. (2024), generalizes the class of Pólya splitting models and does not share the neutral assumptions (not invariant under permutation of species). We propose various areas of investigations, both in probability and in statistics.

2.1 Closure under thinning operator

The closure under addition in Pólya splitting model is equivalent to the closure under the Pólya thinning operator. Thinning operation is a stochastic operation that shrinks a random count variable into a smaller one. This kind of random operation has been intensively studied during the seventies to characterize some count distributions, such as the Poisson distribution using the binomial thinning operator (also named binomial damage model). Then, the closure under thinning operator has been studied in order to define some classes of integer valued autoregressive (INAR) models for count time series. Joe (1996) related the closure under thinning operation to the closure under convolution. More recently, Puig and Valero (2007) characterized the distributions that are closed under the binomial thinning operation. Generalize this result for the class of Pólya thinning operators would allow us to characterize the Pólya splitting distributions (neutral theory) that share the closure under addition.

2.2 Inference of Pólya splitting models

The log-likelihood of a Pólya splitting model is decomposing into two parts according to sum and split respectively. If parameters of both parts are different, then both likelihoods can be maximized separately. The specific models described in Table 1 (in the cases $c = -1$ and $c = 1$) assume a contrast between parameters of sum and split models. This contrast is necessary to obtain the property of closure under addition. Additionally, these models must be considered in a regression context since the environmental variables must be taken into account in ecological application perspectives. Therefore, the formalism of the multivariate link function must be established, as well as the corresponding inference procedure. Until now, the inference of the multinomial ($c = 0$) and Dirichlet multinomial ($c = 1$) regression models have been formalized by Zhang et al. (2017), only for the canonical link function, without considering the sum as random variable.

The candidate should have a strong background in probability and statistics. An interest for counting stochastic processes and statistical modeling would also be welcome, as well as an experience in (R-)programming. An interest in ecological applications and biological processes, will be appreciated.

References

Haegeman, B., Etienne, R.S., 2008. Relaxing the zero-sum assumption in neutral biodiversity theory. Journal of Theoretical Biology 252, 288–294.

Haegeman, B., Etienne, R.S., 2017. A general sampling formula for community structure data. Methods in Ecology and Evolution 8, 1506–1519.

Hubbell, S.P., et al., 2001. The unified neutral theory of biodiversity and biogeography. volume 32. Princeton University Press Princeton.

Joe, H., 1996. Time series models with univariate margins in the convolution-closed infinitely divisible class. Journal of Applied Probability 33, 664–677.

Jones, M., Marchand, É., 2019. Multivariate discrete distributions via sums and shares. Journal of Multivariate Analysis 171, 83–93.

Peyhardi, J., 2023. On quasi pólya thinning operator. Brazilian Journal of Probability and Statistics 37, 643–666.

Peyhardi, J., Fernique, P., 2017. Characterization of convolution splitting graphical models. Statistics & Probability Letters 126, 59–64.

Peyhardi, J., Laroche, F., Mortier, F., 2024. Pólya-splitting distributions as stationary solutions of multivariate birth–death processes under extended neutral theory. Journal of Theoretical Biology 582, 111755.

Puig, P., Valero, J., 2007. Characterization of count data distributions involving additivity and binomial subsampling. Bernoulli , 544–555.

Valiquette, S., Marchand, É., Peyhardi, J., Toulemonde, G., Mortier, F., 2024. Tree polya splitting distributions for multivariate count data. arXiv preprint arXiv:2404.19528 .

Zhang, Y., Zhou, H., Zhou, J., Sun, W., 2017. Regression models for multivariate count data. Journal of Computational and Graphical Statistics 26, 1–13.