Pólya-splitting distributions as stationary solutions of multivariate birth-death processes under extended neutral theory

Jean Peyhardi, Fabien Laroche, Frederic Mortier

Pólya-splitting distributions as stationary solutions of multivariate birth-death processes under extended neutral theory

Jean Peyhardi[1] Fabien Laroche[2]
Frédéric Mortier[3,4]

1. IMAG, University of Montpellier, CNRS, Montpellier, France
2. MR DYNAFOR, INP de Toulouse, INRAE, Auzeville Tolosane, France
3. CIRAD, UPR Forêts et Sociétés, F-34398 Montpellier, France.
4. Environmental Justice Program, Georgetown University, Washington D.C., United States of America

2023

Abstract

Multivariate count distributions are crucial for the inference of ecological processes underpinning biodiversity. In particular, neutral theory provides useful null distributions allowing the evaluation of adaptation or natural selection. In this paper, we build a broader family of multivariate distributions: the Polya-splitting distributions. We show that they emerge naturally as stationary distributions of a multivariate birth-death process. This family of distributions is a consistent extension of non-zero sum neutral models based on a master equation approach. It allows considering both total abundance of the community and relative abundances of species. We emphasize that this family is large enough to encompass various dependence structures among species. We also introduce the strong closure under addition property that can be useful to generate nested multi-level dependence structures. Although all Pólya splitting distributions do not share this property, we provide numerous example verifying it. They include the previously known example with independent species, and also new ones with alternative dependence structures. Overall, we advocate that Polya-splitting distribution should become a part of the classic toolbox for the analysis of multivariate count data in ecology, providing alternative approaches to joint species distribution framework. Comparatively, our approach allows to model dependencies between species at the observation level, while the classical JSDM's model dependencies at the latent process strata.

keywords: Species diversity; Ecological communities; Joint Species Distribution Model; Neutral theory; Multivariate birth-death jump processes; Stationary Distributions.

MSC: 62H05; 60J74; 92D40

1 Introduction

Understanding the processes that shape the biodiversity of ecological communities is a major question in the context of global changes. A large fraction of empirical research about understanding metacommunities relies on pattern-to-process approaches, i.e. detecting the signature of processes through a statistical analysis of the observed distribution of species in space and time. The success of pattern-to-process approaches relies on building an appropriate null hypothesis, where some target process is nullified, and testing whether observations deviate from it.

For 20 years, a family of neutral models, inspired from population genetics, have been proposed as a baseline generating null hypotheses to investigate the effect of species ecological niches. The neutrality assumption consists in assuming that all individuals are ecologically equivalent irrespective of their species, genotypes, etc. hence cancelling any effect associated to ecological niches. One of the most famous example of neutral model is the model introduced by Hubbell (2001). It is a zero-sum game: the total number of individuals in a community is assumed constant, and dead individuals are immediately replaced by offspring of the remaining individuals with equal chance to reproduce for any of them. Neutral model has challenged former pattern-to-process approaches of ecological niches, which were based on permutational apporaches, by showing that neutrality itself could generate non-random structure and thus should be filtered out of patterns using specific models (Bell, 2005; Canard et al., 2012). More precisely, it has been shown that this model yields a local distribution of species abundances within communities that follows a Dirichlet-multinomial distribution (Donnelly et al., 2001; Etienne and Alonso, 2005; Harris et al., 2017). In these studies, the Dirichlet-multinomial distribution has sometimes been called a 'dispersal-limited' multinomial distribution (Etienne and Alonso, 2005).

The Dirichlet-multinomial distribution has several practical interests for pattern-to-process analyses of empirical communities. First, it satisfies useful property denoted weak closure under addition implying in particular that (i) if two species are lumped together, the multivariate distribution of the resulting distribution is still a Dirichlet-multinomial distribution with a natural adaptation of parameters (Laroche et al., 2020); (ii) considering a subgroup of species, the distribution of species abundances conditionnally to the size of the subgroup is also Dirichlet-multinomial, with a natural adaptation of parameters (Laroche et al., 2015). These properties have been used to design several tests in empirical ecological studies. Second, it is readily implemented as a hierarchical process (Harris et al., 2017), hence facilitating the computational aspects or the use of a Bayesian framework for inference purposes.

Importantly, the Dirichlet-multinomial distribution is robust to relaxing the unrealistic zero-sum assumption, by modelling the stochastic dynamics of species abundances within a community as a multivariate jump process. Community size then becomes a random variable fluctuating in time. Using this approach, Haegeman and Etienne (2008) identified a class of neutral models for which decomposing species abundances stationary distribution within a community into a sum distribution (i.e. total number of individuals) and a conditional split distribution (i.e. species labels of individuals) led to a split distribution that is still a

Dirichlet multinomial.

Although those results about Dirichlet-multinomial split have triggered an important and fruitful area of empirical research, they still represent a fraction of the possibilities offered by the neutral theory. Most of the litterature about detecting deviation from neutrality in species abundance data has focused on the split distribution conditionally to the total number of individuals, but to our knowledge only few studies simultaneously discuss the sum and the split (i.e. the full multivariate abundance distribution) as a tool to evaluate community composition (Etienne et al., 2007).However this field has been mostly abandoned in favour to other statistical frameworks based for instance on the multivariate Poisson - lognormal distribution in the context of the joint species distribution model (JSDM) (Aïtchison and Ho, 1989; Warton et al., 2015; Ovaskainen and Abrego, 2020). In addition, studies focusing on the split distribution have mostly focused on formulations of the neutrality assumption and ancillary hypotheses that necessarily lead to Dirichlet-multinomial distributions. If these assumptions are relaxed or modified, we expect that new split distributions can be obtained in models that still arguably remain neutral.

Our aim here is to show how relaxing some assumptions of neutral models can generalize the sum-split decomposition with Dirichlet-multinomial split to the more general Polya-splitting distributions family. We show that our framework covers classical distributions but also lead to new ones. Hence our work promotes a unified neutral-based statistical framework able to tackle the full multivariate abundance distribution of species, thus making an interesting mechanistic alternative to current phenomenological JSDMs framework. We also present a new general property denoted the strong closure under addition. This property allows in particular proposing approaches based on recursive application of splitting distributions to generate communities mixing dependent or independent species or group of species simultaneously

Section 2 describes the family of multivariate Pólya splitting distributions. We specifically focus on nine examples of such distributions sharing the property of strong closure under addition. Section 3 shows that these distributions are stationary solutions of the master equation under a specific parametric hypothesis on the ratio between birth and death rates. Section 4 shows that this parametric assumption corresponds to a mildly extended version of the neutral theory of biodiversity where the immigration rate of a species can depend on its local abundance following a relationship identical across all species.

2 Polya splitting distributions

The first subsection recall the definition of multivariate Pólya distributions as urn models with n random drawing. Then, assuming that n is a random number, the second subsection presents the larger family of Pólya splitting distributions, introduced by Peyhardi and Fernique (2017) and more generally studied by Peyhardi et al. (2021); Peyhardi (2023). Nine examples of Pólya splitting distributions with remarkable properties are presented, which are repeatedly referred to in the rest of the article.

2.1 Multivariate Pólya distributions

The Pólya urn model is generally presented in terms of n random drawings of balls from an urn, that initially contains $\theta_j \in \mathbb{N}^*$ balls of the j^{th} color. One ball is drawn at random and then replaced with $c \in \mathbb{Z}$ additional balls of the same color. A negative value for c means that balls are removed from the urn. This procedure is repeated n times and focus is made on the count $\boldsymbol{N} = (N_1, \ldots, N_J)$ of drawn balls for the $J \geq 2$ different colors. Let $|\boldsymbol{N}| = \sum_{j=1}^{J} N_j$ denotes the sum of the vector $\boldsymbol{N}$ and $\Delta_n = \left\{\boldsymbol{n} \in \mathbb{N}^J : |\boldsymbol{n}| = n\right\}$ (resp. $\blacktriangle_n = \left\{\boldsymbol{n} \in \mathbb{N}^J : |\boldsymbol{n}| \leq n\right\}$) the discrete simplex (resp. the discrete corner of the hypercube). The multivariate count distribution for $\boldsymbol{N}$ is known as the multivariate Pólya distribution and will be denoted by $\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta})$. Its probability mass function (pmf) is given by

$$P_{|\boldsymbol{N}|=n}\left(\boldsymbol{N} = \boldsymbol{n}\right) = \frac{\displaystyle\prod_{j=1}^{J}\prod_{k=0}^{n_j-1} r_{\theta_j}^{[c]}(k_j)}{\displaystyle\prod_{k=0}^{n-1} r_{|\boldsymbol{\theta}|}^{[c]}(k)},$$

where $r_{\theta}^{[c]}(k) = \frac{\theta+ck}{k+1}\mathbb{1}_{\theta+ck\geq 0}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J) \in \Theta_c^J$. The indicator function ensures that $r_{\theta}^{[c]}(k) \geq 0$ even if $c < 0$. Let us define $R_{\theta}^{[c]}(n) = \prod_{k=0}^{n-1} r_{\theta}^{[c]}(k)$, then the pmf becomes

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

The multivariate Pólya distribution turns out to be the multivariate hyergeometric distribution when $c = -1$, the multinomial distribution when $c = 0$ and the multivariate negative hyergeometric distribution when $c = 1$. Reasoning by equivalence on the pmf, it can be shown that these three distributions are the representative elements of their equivalence classes: $\{\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) : c < 0\}$, $\{\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) : c = 0\}$ and $\{\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) : c > 0\}$. Therefore, in the following, a focus will be made only on the three cases $c \in \{-1, 0, 1\}$. Let us note that the pmf for the last two cases can be extended for continuous values of $\theta_j \in \mathbb{R}_+^*$ for $j = 1, \ldots, J$. The multivariate negative hyergeometric distribution is thereby extended to the Dirichlet multinomial distribution. In summary, the three cases $c \in \{-1, 0, 1\}$ respectively correspond to the

- multivariate hypergeometric distribution, denoted by $\mathcal{H}_{\Delta_n}(\boldsymbol{\theta})$ with $\Theta_{-1} = \mathbb{N}^*$ and $R_{\theta}^{[-1]}(n) = \binom{\theta}{n}$,
- multinomial distribution, denoted by $\mathcal{M}_{\Delta_n}(\boldsymbol{\pi})$ where $\boldsymbol{\pi} = \boldsymbol{\theta}/|\boldsymbol{\theta}|$, $\Theta_0 = \mathbb{R}_+^*$ and $R_{\theta}^{[0]}(n) = \theta^n$,
- Dirichlet-multinomial distribution, denoted by $\mathcal{DM}_{\Delta_n}(\boldsymbol{\theta})$ where $\Theta_1 = \mathbb{R}_+^*$ and $R_{\theta}^{[1]}(n) = \binom{n+\theta-1}{n}$,

The pmf of these three distributions are presented in Table A.2 of Appendix A. The support is Δ_n when $c = 0$ or $c = 1$ and is $\Delta_n \cap \blacksquare_{\boldsymbol{\theta}}$ when $c = -1$, i.e., the intersection between the simplex Δ_n and the hyper-rectangle $\blacksquare_{\boldsymbol{\theta}} = \{\boldsymbol{n} \in \mathbb{N}^J : n_1 \leq \theta_1, \ldots, n_J \leq \theta_J\}$. It should be noted that some authors refers to the Dirichlet multinomial distribution as the multivariate Pólya distribution. All along the paper, the multivariate Pólya distribution will refer to the general case that encompasses the three cases $c \in \{-1, 0, 1\}$.

2.2 Specific Pólya splitting distributions with remarkable properties

Pólya distributions cannot be considered as a *sensu stricto* multivariate distribution. Indeed, the sum of the random vector $\boldsymbol{N}$ is fixed and only $J - 1$ elements over J are free. This kind of distribution, supported on Δ_n, is said to be singular. It is possible to define a non-singular version, supported on $\blacktriangle_n$. The vector $\boldsymbol{N}$ is said to follow a non-singular Pólya distribution, denoted by $\mathcal{P}^{[c]}_{\blacktriangle_n}(\boldsymbol{\theta}, \gamma)$ with additional parameter $\gamma \in \Theta_c$, if the completed vector $(\boldsymbol{N}, m - |\boldsymbol{N}|)$ follows the singular version $\mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}, \gamma)$. However, the support of this extension remains bounded (see Table A.2 of Appendix A for details about the non-singular version of Pólya distributions).

Another way to relax the fixed sum assumption, is considering the sum $|\boldsymbol{N}|$ as a random variable. The sum $|\boldsymbol{N}|$ then follows an univariate count distribution $\mathcal{L}(\psi)$ and the vector $\boldsymbol{N} = (N_1, \ldots, N_J)$ given the sum $|\boldsymbol{N}| = n$ follows a multivariate Pólya distribution $\mathcal{P}^{[c]}_{\Delta_n}(\theta_1, \ldots, \theta_J)$. We thus obtain a Polya splitting distribution, that can be viewed as a compound distribution denoted as follows:

$$\boldsymbol{N} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\psi),$$

where ψ is an univariate or multivariate set of unknown parameters (e.g. univariate for the Poisson distribution and bivariate for the negative binomial distribution). To be more explicit the pmf of the Pólya splitting distribution is given by $P(\boldsymbol{N} = \boldsymbol{n}) = P_{|\boldsymbol{N}|=|\boldsymbol{n}|}(\boldsymbol{N} = \boldsymbol{n})P(|\boldsymbol{N}| = |\boldsymbol{n}|)$. According to $\mathcal{L}(\psi)$, several multivariate distributions can be defined sharing interesting properties.

The *weak closure under addition* The *weak closure under addition* is the stability under convolution of $R^{[c]}_{\theta}$: for all $\theta, \theta' \in \Theta_c$

$$R^{[c]}_{\theta} * R^{[c]}_{\theta'} = R^{[c]}_{\theta+\theta'}, \tag{1}$$

where $*$ denotes the discrete convolution. This identity plays a central role among the family of Polya splitting distributions and corresponds to the Vandermonde's identity, the Newton's binomial theorem and the Hagen-Rothe's identity when $c = -1, 0$ and 1 respectively. It could be shown that equation (1) implies the stability of the multivariate distribution under marginalization; see Theorem 1 of Peyhardi et al. (2021) for details. Let focus on two such stability properties.

(i) if two species are lumped together, the multivariate distribution of the resulting distribution is still a Pólya splitting distribution with a natural adaptation of parameters

(Laroche et al., 2020), i.e., we have

$$(N_1 + N_2, N_3, \ldots, N_J) \sim \mathcal{P}^{[c]}_{\Delta_n}(\theta_1 + \theta_2, \theta_3, \ldots, \theta_J) \underset{n}{\wedge} \mathcal{L}(\psi),$$

(ii) considering a subgroup of species, the distribution of species abundances conditionally to the size of the subgroup is also Pólya distribution with a natural adaptation of parameters (Laroche et al., 2015), i.e., we have

$$(N_1, N_2, N_3) | N_1 + N_2 + N_3 = n \sim \mathcal{P}^{[c]}_{\Delta_n}(\theta_1, \theta_2, \theta_3).$$

It should be noted that all Pólya splitting distributions share this property, i.e., it holds for any sum distribution $\mathcal{L}(\psi)$.

The *strong closure under addition* A Pólya splitting distribution is said to be strongly closed under addition if the sum distribution and all the marginal distributions belong to the same family of parametric distributions, i.e., if we have

$$\forall j \in \{1, \ldots, J\} \; \exists \psi_j \; : \; N_j \sim \mathcal{L}(\psi_j).$$

While weak property is share by all Pólya splitting distributions, strong property holds for only specific sum distribution see Peyhardi (2023) for details about this closure property. In the following, we propose nine sum distributions verifying strong closure property. The first three distributions naturally extend singular Pólya distributions to their non-singular version. The three others leads to independence splitting distributions. Finally, the three last distributions allows the generalization of classical univariate count processes to their multivariate version.

Canonical cases

In the canonical case, the sum distribution is defined as the univariate version of the non-singular Pólya distribution $\mathcal{P}^{[c]}_{\blacktriangle_m}(\boldsymbol{\theta}, \gamma)$. The sum distribution is then denoted by $\mathcal{P}^{[c]}_{m}(\theta, \gamma)$; see Table A.1 for details about its pmf and support. Theorem 4 of Peyhardi et al. (2021) showed that we have the following distribution identity

$$\mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{P}^{[c]}_{m}(|\boldsymbol{\theta}|, \gamma) = \mathcal{P}^{[c]}_{\blacktriangle_m}(\boldsymbol{\theta}, \gamma).$$

See Table 1 to write this identity in the three cases $c \in \{-1, 0, 1\}$. It could be shown that marginals belongs to the same family of distribution, more precisely $N_j \sim \mathcal{P}^{[c]}_{m}(\theta_j, |\boldsymbol{\theta}_{-j}| + \gamma)$ (the strong closure under addition holds). The non-singular version has the advantage that its support $\blacktriangle_n$ has a dimension equal to J (whereas the support of the singular version Δ_n has a dimension equal to $J - 1$). Therefore the J variables $N_1, \ldots, N_J$ are free (not related by linear function) even if they are not independent. The graphical model of independence of such a distribution is complete (Peyhardi, 2023). In summary for the canonical case, the variables are free and not independent but the support is bounded.

Independence cases

There exists a sum distribution $\mathcal{L}^*$ such that all the variables $N_1, \ldots, N_J$ are mutually independent. Moreover, all the margins belong to the same family $\mathcal{L}^*$, i.e. the strong closure under addition holds. It could be shown that the pmf of this distribution is

$$P(|\boldsymbol{N}| = n) = \frac{R^{[c]}_{|\boldsymbol{\theta}|}(n)\alpha^n}{\sum_{m=0}^{\infty} R^{[c]}_{|\boldsymbol{\theta}|}(m)\alpha^m}, \tag{2}$$

for some $\alpha > 0$ and the marginals are given by

$$P(N_j = n) = \frac{R^{[c]}_{\theta_j}(n)\alpha^n}{\sum_{m=0}^{\infty} R^{[c]}_{\theta_j}(m)\alpha^m}. \tag{3}$$

The distribution $\mathcal{L}^*$ belongs to the family of power series distributions. It turns out to be the binomial distribution ($c = -1$), the Poisson distribution ($c = 0$) and the negative binomial distribution ($c = 1$) respectively; see Table 1 for details.

Dependent non-canonical cases

The last line of Table 1 presents three Pólya splitting distributions that share the strong closure under addition without independence assumption and different of the canonical case (demonstrations are given in Appendix B). Those distributions are obtained from the independent case assuming the parameter α (see eq. (2)) is a random variable. For instance when $c = -1$, the sum distribution is a binomial distribution compound by a beta distribution, i.e., a beta binomial distribution. According to (3), it is easily seen that marginals also follow beta binomial distributions. This particular distribution can be viewed as a new multivariate extension of the beta binomial distribution different from the more usual Dirichlet multinomial (non singular version). It is supported on the hyper-rectangle $\blacksquare_{\boldsymbol{\theta}}$. For $c = 0$, the Poisson distribution is coumpound by a gamma distribution to obtain a negative binomial distribution. The corresponding Pólya splitting distribution turns out to be the natural multivariate extension, i.e., the negative multinomial distribution. Finally, for $c = 1$, the negative binomial is compound by a beta distribution to obtain a negative beta binomial distribution, also known as the univariate generalized Waring distribution (Irwin, 1968). The corresponding Pólya splitting distribution turns out to be its natural multivariate extension, known as the multivariate generalized Waring distribution (MGWD) (Xekalaki, 1986). This dependent and non-canonical case has been recently formalized by Peyhardi (2023) as the inverse Pólya distribution.

3 Stationary distributions of multivariate birth-death processes

We show that the class of Pólya splitting distributions introduced in former section exactly corresponds to the stationary distributions of multivariate birth-death processes under specific parametric assumptions on jumping rates. We further provide more precise conditions

	Hypergeometric splitting distributions $c=-1$	Multinomial splitting distributions $c=0$	Dirichlet multinomial splitting distributions $c=1$
Canonical cases	$\mathcal{H}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{H}_m(\lvert\boldsymbol{\theta}\rvert,\gamma)$ $=$ $\mathcal{H}_{\blacktriangle_m}(\boldsymbol{\theta},\gamma)$ $\boldsymbol{\theta}\in\mathbb{N}^{*J}$, $\gamma\in\mathbb{N}^*$, $m\in\mathbb{N}^*$, $m\leq\lvert\boldsymbol{\theta}\rvert+\gamma$ (support $=(\blacktriangle_m\setminus\blacktriangle_{m-\gamma})\cap\blacksquare_{\boldsymbol{\theta}}$)	$\mathcal{M}_{\Delta_n}(\boldsymbol{\pi}) \underset{n}{\wedge} \mathcal{B}_m(p)$ $=$ $\mathcal{M}_{\blacktriangle_m}(p\cdot\boldsymbol{\pi})$ $\boldsymbol{\pi}\in\Delta$, $m\in\mathbb{N}^*$ (support $=\blacktriangle_m$)	$\mathcal{DM}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \beta\mathcal{B}_m(\lvert\boldsymbol{\theta}\rvert,\gamma)$ $=$ $\mathcal{DM}_{\blacktriangle_m}(\boldsymbol{\theta},\gamma)$ $\boldsymbol{\theta}\in\mathbb{R}_+^{*J}$, $\gamma\in\mathbb{R}_+^*$, $m\in\mathbb{N}^*$ (support $=\blacktriangle_m$)
Independent cases	$\mathcal{H}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{B}_{\lvert\boldsymbol{\theta}\rvert}(p)$ $=$ $\bigotimes_{j=1}^{J}\mathcal{B}_{\theta_j}(p)$ $\boldsymbol{\theta}\in\mathbb{N}^{*J}$, $p\in(0,1)$ (support $=\blacksquare_{\boldsymbol{\theta}}$)	$\mathcal{M}_{\Delta_n}(\boldsymbol{\pi}) \underset{n}{\wedge} \mathcal{P}(\lambda)$ $=$ $\bigotimes_{j=1}^{J}\mathcal{P}(\pi_j\lambda)$ $\boldsymbol{\pi}\in\Delta$, $\lambda\in\mathbb{R}_+^*$ (support $=\mathbb{N}^J$)	$\mathcal{DM}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{NB}(\lvert\boldsymbol{\theta}\rvert,p)$ $=$ $\bigotimes_{j=1}^{J}\mathcal{NB}(\theta_j,p)$ $\boldsymbol{\theta}\in\mathbb{R}_+^{*J}$, $p\in(0,1)$ (support $=\mathbb{N}^J$)
Dependent non-canonical cases	$\mathcal{H}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \beta\mathcal{B}_{\lvert\boldsymbol{\theta}\rvert}(a,b)$ $=$ $\mathcal{M}\beta_{\blacksquare_{\boldsymbol{\theta}}}(a,b)$ $\boldsymbol{\theta}\in\mathbb{N}^{*J}$, $a\in\mathbb{R}_+^*$, $b\in\mathbb{R}_+^*$ (support $=\blacksquare_{\boldsymbol{\theta}}$)	$\mathcal{M}_{\Delta_n}(\boldsymbol{\pi}) \underset{n}{\wedge} \mathcal{NB}(a,p)$ $=$ $\mathcal{NM}(a,p\cdot\boldsymbol{\pi})$ $\boldsymbol{\pi}\in\Delta$, $a\in\mathbb{R}_+^*$, $b\in\mathbb{R}_+^*$ (support $=\mathbb{N}^J$)	$\mathcal{DM}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \beta\mathcal{NB}(\lvert\boldsymbol{\theta}\rvert,a,b)$ $=$ $\mathrm{MGWD}(b,\boldsymbol{\theta},a)$ $\boldsymbol{\theta}\in\mathbb{R}_+^{*J}$, $a\in\mathbb{R}_+^*$, $b\in\mathbb{R}_+^*$ (support $=\mathbb{N}^J$)

Table 1: Nine Pólya splitting distributions that share the strong closure under addition. The notation $\boldsymbol{N}\sim\bigotimes_{j=1}^{J}\mathcal{L}_j$ means that each N_j follows the distribution $\mathcal{L}_j$ independently of all other $j'\neq j$.

on rates that lead to the nine Pólya splitting distributions previously described (see Table 1). Let us start with the univariate case ($J=1$) to recall classical results that will be useful to explicitly describe the stationary distribution of the sum in the multivariate case ($J\geq 2$).

3.1 The univariate case

Let $N(t)$ denote an univariate birth/death process with $q^+(n)$ (resp. $q^-(n)$) denoting the birth rate (resp. death rate) for a population of size n. Let $p_n=P(N=n)$ denote the pmf of the stationary distribution. It can be shown, by induction on n, that solutions of the master equation at stationary state are solutions of the detailed balance equation:

$$q^-(n+1)p_{n+1}=q^+(n)p_n$$

Remark that for every count distribution $(p_n)_{n\in\mathbb{N}}$, it is possible to find two sequences of birth and death rates such that the detailed balance holds. Moreover, assuming that the death rate is positive, i.e., $q^-(n)>0$ for all $n\geq 1$, then the detailed balance becomes

$$p_{n+1}=p_n q(n),$$

where $q(n) = q^+(n)/q^-(n+1)$. The support of such a distribution is necessary of the connected form $\{0, \ldots, m\}$ with $m \in \mathbb{N} \cup \{\infty\}$ and we have:

$$p_n = \frac{Q_n}{\sum_{k \geq 0} Q_k},$$

where $Q_n = \prod_{k=0}^{n-1} q(k)$ for $n \geq 1$ and $Q_0 = 1$. Specific stationary distribution are obtain according to specific parametric assumption on $q(n)$. For instance if $q(n) = \frac{1}{n+1}\alpha$ for some $\alpha \in \mathbb{R}_+^*$ then the stationary distribution is a Poisson distribution with parameter α. See Appendix C for several examples of parametric assumption on $q(n)$ that lead to usual univariate distributions (e.g. binomial, negative binomial).

3.2 The multivariate case

Here we now describe the multivariate jump process $\boldsymbol{N}(t) = \{N_1(t), \ldots, N_J(t)\}$ depicting species abundances within a community. We then focus on sufficient conditions on jumping rates to ensure the existence of a stationary distribution with detailed balance. In this specific case, it is straightforward to derive a closed form of the stationary distribution. In the following let $p_{\boldsymbol{n}}(t) = P\{\boldsymbol{N}(t) = \boldsymbol{n}\}$ denote the pmf at time t and $p_{\boldsymbol{n}} = P(\boldsymbol{N} = \boldsymbol{n})$ denote the pmf at stationary state.

The master equation

$$\frac{\partial p_{\boldsymbol{n}}(t)}{\partial t} = \sum_{j=1}^{J} p_{\boldsymbol{n}-\boldsymbol{e}_j}(t) q_j^-(\boldsymbol{n}-\boldsymbol{e}_j) + p_{\boldsymbol{n}+\boldsymbol{e}_j}(t) q_j^+(\boldsymbol{n}+\boldsymbol{e}_j) - p_{\boldsymbol{n}+\boldsymbol{e}_j}(t)\{q_j^-(\boldsymbol{n}) + q_j^+(\boldsymbol{n})\} \quad (4)$$

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 $\boldsymbol{n} + \boldsymbol{e}_j$) and $\boldsymbol{e_i}$ is a vector of size J where the elements are all equal to zero except the i^{th} equal to one. It is usual to assume that $q_j^-(\boldsymbol{n}) > 0$ for all $\boldsymbol{n} \in \mathbb{N}^J$ such that $n_j > 0$ (i.e., any individual is mortal). Moreover it is assumed that $q_j^+(\boldsymbol{0}) > 0$ for all $j = 1, \ldots, J$ (where $\boldsymbol{0} = (0, \ldots, 0)$ denotes the null vector) in order to avoid the case of non observed species (i.e., $P(N_j = 0) = 1$).

The Detailed balance equation is given, for all $j \in \{1, \ldots, J\}$ and all $\boldsymbol{n} \in \mathbb{N}^J$, by

$$p_{\boldsymbol{n}+\boldsymbol{e_j}} q_j^-(\boldsymbol{n}+\boldsymbol{e_j}) = p_{\boldsymbol{n}}\, q_j^+(\boldsymbol{n}). \quad (5)$$

Since it is assumed that $q_j^-(\boldsymbol{n}) > 0$ for all $\boldsymbol{n} \in \mathbb{N}^J$ such that $n_j > 0$ then the detailed balance becomes

$$p_{\boldsymbol{n}+\boldsymbol{e_j}} = p_{\boldsymbol{n}}\, q_j(\boldsymbol{n}) \quad (6)$$

where $q_j(\boldsymbol{n}) = q_j^+(\boldsymbol{n})/q_j^-(\boldsymbol{n}+\boldsymbol{e_j})$. The idea is then to recursively use the equality (6) in order to express $p_{\boldsymbol{n}}$ according to $p_{\boldsymbol{0}}$ and thus derive a closed analytical formula for the stationary distribution pmf of the multivariate jump process.

Necessary and sufficient conditions for detailed balance Note that the detailed balance holds if and only if the product of the quantitites q_j along every increasing path between $\mathbf{0}$ and $\boldsymbol{n}$ is the same. This is equivalent to assume that this product is constant along every path between $\boldsymbol{n}$ and $\boldsymbol{n}+\boldsymbol{e_i}+\boldsymbol{e_j}$ with $i \neq j$. There are only two such path: $\boldsymbol{n} \to \boldsymbol{n}+\boldsymbol{e_i} \to \boldsymbol{n}+\boldsymbol{e_i}+\boldsymbol{e_j}$ and $\boldsymbol{n} \dashrightarrow \boldsymbol{n}+\boldsymbol{e_j} \dashrightarrow \boldsymbol{n}+\boldsymbol{e_j}+\boldsymbol{e_i}$. For the first path the detailed balance equation gives $p_{\boldsymbol{n}+\boldsymbol{e_i}+\boldsymbol{e_j}} = p_{\boldsymbol{n}}\ q_i(\boldsymbol{n})\ q_j(\boldsymbol{n}+\boldsymbol{e_i})$. For the second path the detailed balance equation gives $p_{\boldsymbol{n}+\boldsymbol{e_i}+\boldsymbol{e_j}} = p_{\boldsymbol{n}}\ q_j(\boldsymbol{n})\ q_i(\boldsymbol{n}+\boldsymbol{e_j})$. Therefore a necessary and sufficient condition to the existence of a solution is the equality

$$q_i(\boldsymbol{n})\ q_j(\boldsymbol{n}+\boldsymbol{e_i}) = q_j(\boldsymbol{n})\ q_i(\boldsymbol{n}+\boldsymbol{e_j}), \tag{7}$$

for all $i \neq j \in \{1,\dots,J\}$ and all $\boldsymbol{n} \in \mathbb{N}^J$ such that $p_{\boldsymbol{n}} \neq 0$. Equation (7) corresponds to the Kolmogorov's criterion in the case of a multivariate birth-death process with positive death rates. This criterion is a necessary and sufficient condition for the reversibility of the process.

Parametric assumption on birth and death rates Assume that there exists some parameters $c \in \{-1,0,1\}$ and $\boldsymbol{\theta} = (\theta_1,\dots,\theta_J) \in \Theta_c^J$ and two non-negative functions s^+ and s^- such that the birth and death rates have the following form

$$\begin{array}{l} q_j^+(\boldsymbol{n}) = s^+(|\boldsymbol{n}|)(\theta_j + cn_j)\mathbb{1}_{\theta_j+cn_j\geq 0}, \\ q_j^-(\boldsymbol{n}) = s^-(|\boldsymbol{n}|)n_j. \end{array} \tag{8}$$

The birth-deah rate thus becomes

$$q_j(\boldsymbol{n}) = s(|\boldsymbol{n}|)\ r_{\theta_j}^{[c]}(n_j). \tag{9}$$

where $s(n) = \frac{s^+(n)}{s^-(n+1)}$ for all $n \in \mathbb{N}$ and $r_\theta^{[c]}(n) = \frac{\theta+cn}{n+1}\mathbb{1}_{\theta+cn\geq 0}$. It is easily seen that this parametric assumption (9) respects the Kolmogorov's criterion (7) and thus the detailed balance equation (5). In order to obtain a well defined stationary distribution we add the following assumption on $s(n)$:

$$\sum_{n\geq 0} R_{|\boldsymbol{\theta}|}^{[c]}(n) \prod_{k=0}^{n-1} s(k) < \infty. \tag{10}$$

Theorem 1 *Assume that the hypothesis* (9) *and* (10) *hold then*

- *the stationary distribution for* $\boldsymbol{N} = (N_1,\dots,N_J)$ *is the Polya splitting distribution* $\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}$
- $\mathcal{L}$ *is the stationary distribution of a univariate process with birth/death ratio equal to* $q(n) = s(n)r_{|\boldsymbol{\theta}|}^{[c]}(n)$ *(its support is necessary of the form* $\{0,\dots,m\}$ *where* $m \in \mathbb{N}\cup\{\infty\}$*).*

Proof: Under assumption (9), using recursively (6), we obtain the pmf of the stationary distribution as follows:

$$p_{\boldsymbol{n}} = p_{\mathbf{0}} \prod_{m=0}^{|\boldsymbol{n}|-1} s(m) \prod_{j=1}^{J} \prod_{k=0}^{n_j-1} r_{\theta_j}^{[c]}(k).$$

Using notations of Section 2 we obtain

$$p_{\boldsymbol{n}} = p_{\mathbf{0}} \prod_{m=0}^{|\boldsymbol{n}|-1} s(m) \prod_{j=1}^{J} R_{\theta_j}^{[c]}(n_j).$$

Since the pmf is written as a recurrent product, it can be remarked that the support of the stationary distribution is exactly the connexe $\blacktriangle_m$ when $c = 0$ or $c = 1$ and the connexe $\blacktriangle_m \cap \blacksquare_{\boldsymbol{\theta}}$ when $c = -1$, where m is the smaller integer such that $s(m) = 0$. Note that if $m = +\infty$ then the support becomes $\mathbb{N}^J$ when $c = 0$ or $c = 1$ and $\blacksquare_{\boldsymbol{\theta}}$ when $c = -1$. Otherwise, We know that the probability of $\boldsymbol{n}$ can be conditioned by the sum as follows:

$$p_{\boldsymbol{n}} = P(|\boldsymbol{N}| = |\boldsymbol{n}|)\ P_{|\boldsymbol{N}|=|\boldsymbol{n}|}(\boldsymbol{N} = \boldsymbol{n}).$$

By identifiability between the two previous equalities (on the support) we obtain the sum distribution and the split distribution (repartition into components given the sum). More precisely we have

$$P_{|\boldsymbol{N}|=n}(\boldsymbol{N} = \boldsymbol{n}) = \frac{\prod_{j=1}^{J} R_{\theta_j}^{[c]}(n_j)}{C(n)},$$

for all $n \in \mathbb{N}$, where $C(n)$ is the normalizing constant

$$C(n) = \sum_{\boldsymbol{n} \in \Delta_n} \prod_{j=1}^{J} R_{\theta_j}^{[c]}(n_j),$$

and

$$P(|\boldsymbol{N}| = n) = \frac{C(n) \prod_{k=0}^{n-1} s(k)}{\sum_{m \geq 0} C(m) \prod_{k=0}^{m-1} s(k)}.$$

Note that $C(n)$ is positive since the identification is made on the support. Now remark that $C(n)$ turns out to be the convolution $C = R_{\theta_1}^{[c]} * \cdots * R_{\theta_J}^{[c]}$. According to the Newton's binomial theorem, for $c = 0$, (respectively the Vandermonde's identity for $c = -1$ and the Hagen-Rothe identity for $c = 1$) we have

$$R_{\theta_1}^{[c]} * \cdots * R_{\theta_J}^{[c]} = R_{|\boldsymbol{\theta}|}^{[c]},$$

and thus the pmf of $\boldsymbol{N}$ given the sum $|\boldsymbol{N}| = n$ is

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

i.e., given $|\boldsymbol{N}| = n$ we have $\boldsymbol{N} \sim \mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta})$ (multivariate Polya distribution). The sum distribution is now given by

$$P(|\boldsymbol{N}| = n) = \frac{R_{|\boldsymbol{\theta}|}^{[c]}(n) \prod_{k=0}^{n-1} s(k)}{\sum_{m \geq 0} R_{|\boldsymbol{\theta}|}^{[c]}(m) \prod_{k=0}^{m-1} s(k)}. \tag{11}$$

This is a proper distribution according to assumption (10). Moreover, by definition we have $R_{|\boldsymbol{\theta}|}^{[c]}(n) = \prod_{k=0}^{n-1} r_{|\boldsymbol{\theta}|}^{[c]}(k)$ (with convention $R_{|\boldsymbol{\theta}|}^{[c]}(0) = 1$) and so the sum distribution can be viewed as the stationary distribution of an univariate process whose the ratio of birth/death is $q(n) = s(n) r_{|\boldsymbol{\theta}|}(n)$.

Parametric hypothesis on $s(n)$

For any univariate count distribution $\mathcal{L}$ with a support of the form $\{0,\dots,m\}$ where $m \in \mathbb{N}\cup\{\infty\}$, there exists a birth/death rate $q(n)$ (and thus a function $s(n)$) such that the corresponding birth-death process converges toward the stationary distribution $\mathcal{L}$. Therefore, for any Pólya splitting distribution $\mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}$ (with a sum support of the form $\{0,\dots,m\}$ where $m \in \mathbb{N}\cup\{\infty\}$) we are able to write the birth and death rates (through $r_\theta(n)$ and $s(n)$) such that the multivariate jump process converges toward the stationary distribution $\mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}$. Let us illustrate this fact with nine parametric assumptions on $s(n)$ that lead to the nine Pólya splitting distributions of Table 1. According to Theorem 1 the sum is a univariate birth-death process driven by the birth/death rate $q(n) = s(n)r^{[c]}_{|\boldsymbol{\theta}|}(n)$. Appendix C details the parametric form for $q(n)$ that leads to specific univariate distributions (binomial, Poisson, etc ...). By identification it is possible to find the parametric form of $s(n)$ in each case $c \in \{-1,0,1\}$ and deduce the sum distribution as compiled in table 2. It is also possible to find the parametric form of $s(n)$ in a general way, i.e., for any $c \in \{-1,0,1\}$; see Appendix D for details about the canonical cases.

Setting apart the canonical case with $c = 1$ and $\gamma < 1$, which will be discussed later, we observe clear qualitative differences among the nine examples. Each case corresponds to a particular variation profile of $s(n)$. The function s decreases in the canonical case, is constant in the independent case and increases in the dependent non-canonical case (see Appendix F for details). Moreover, as c increases in $\{-1,0,1\}$, the convexity of $s(n)$ shows consistent changes which do not depend on the considered case. When $c = -1$, $s(n)$ is convex, when $c = 0$, $s(n)$ is linear and, when $c = 1$, $s(n)$ is concave.

The exception to above pattern is the canonical case with $c = 1$ when $\gamma < 1$. In this case, $s(n)$ is increasing and convex, thus resembling the dependent non-canonical case with $c = -1$. This exception is obtained when the sum follows the beta binomial distribution $\beta\mathcal{B}_m(|\boldsymbol{\theta}|,\gamma)$ with $\gamma < 1$ inducing a peak at $n = m$, hence promoting saturation of community size at m.

	$c=-1$	$c=0$	$c=1$
Canonical cases	$s(n) = \dfrac{m-n}{\gamma-m+n+1}\mathbb{1}_{m-\gamma\le n<m}$ $\gamma\in\mathbb{N}^*,\ m\in\mathbb{N}^*,\ m\le\lvert\boldsymbol{\theta}\rvert+\gamma$	$s(n) = \dfrac{m-n}{\gamma}\mathbb{1}_{n<m}$ $\gamma\in\mathbb{R}_+^*,\ m\in\mathbb{N}^*$	$s(n) = \dfrac{m-n}{\gamma+m-n-1}\mathbb{1}_{n<m}$ $\gamma\in\mathbb{R}_+^*,\ m\in\mathbb{N}^*$
Independent cases	$s(n)=\alpha$ $\alpha\in\mathbb{R}_+^*$	$s(n)=\alpha$ $\alpha\in\mathbb{R}_+^*$	$s(n)=\alpha$ $\alpha\in(0,1)$
Dependent non-canonical cases	$s(n) = \dfrac{a+n}{\lvert\boldsymbol{\theta}\rvert+b-n-1}\mathbb{1}_{n<\lvert\boldsymbol{\theta}\rvert}$ $a\in\mathbb{R}_+^*$ and $b\in\mathbb{R}_+^*$	$s(n) = \dfrac{a+n}{\lvert\boldsymbol{\theta}\rvert+b}$ $a\in\mathbb{R}_+^*$ and $b\in\mathbb{R}_+^*$	$s(n) = \dfrac{b+n}{\lvert\boldsymbol{\theta}\rvert+a+b+n}$ $a\in\mathbb{R}_+^*$ and $b\in\mathbb{R}_+^*$

Table 2: Parametric hypothesis on $s(n)$ that lead to the nine Pólya splitting distributions of Table 1

4 Biological interpretations of the Pólya splitting distributions

From a biological perspective, the birth and death rates of the multivariate jump process defined by the master equation (see eq. (4)) are generally assumed to have the following form:

$$\begin{aligned} q_j^+(\boldsymbol{n}) &= m_j(\boldsymbol{n}) + n_j b_j(\boldsymbol{n}) \\ q_j^-(\boldsymbol{n}) &= n_j d_j(\boldsymbol{n}) \end{aligned}$$

where $m_j(\boldsymbol{n})$ is the immigration rate of species j, $b_j(\boldsymbol{n})$ is the per-capita local reproduction rate of species j and $d_j(\boldsymbol{n})$ is the per-capita local death-or-emigration rate of species j. Classical neutral models make two additional assumptions. The first corresponds to the strong neutrality:

$$\begin{aligned} q_j^+(\boldsymbol{n}) &= m(|\boldsymbol{n}|)\pi_j + n_j b(|\boldsymbol{n}|) \\ q_j^-(\boldsymbol{n}) &= n_j d(|\boldsymbol{n}|) \end{aligned} \tag{12}$$

The second assumption is detailed balance. Given the strong neutrality assumption and discarding degenerated cases $m(0) = 0$ or $m(1) = 0$, detailed balance occurs if and only if there exists a constant $\tilde{I} \geq 0$ such that

$$b(n) = \tilde{I} m(n) \tag{13}$$

for all $n \in \mathbb{N}^*$ (see Appendix E). The parameter $I = 1/\tilde{I}$ his known as the effective number of migrants (Etienne and Olff, 2004). Here we focus on a generalization of equations (12), considering new expressions of birth and death rates as follows:

$$\begin{aligned} q_j^+(\boldsymbol{n}) &= \left[m(|\boldsymbol{n}|) \left(\pi_j - \tilde{K} n_j \right) + n_j b(|\boldsymbol{n}|) \right] \mathbb{1}_{\pi_j - \tilde{K} n_j \geq 0} \\ q_j^-(\boldsymbol{n}) &= n_j d(|\boldsymbol{n}|) \end{aligned} \tag{14}$$

with $\tilde{K} \in \mathbb{R}_+$. The master equation with assumption (14) can be called a neutral model with density-dependent immigration, because the rate of immigration of species j is now $m(|\boldsymbol{n}|)(\pi_j - \tilde{K} n_j)$, which depends on species j and on the local population size n_j. Assuming $\tilde{K} = 0$ leads to the usual equations (12). We still assume (13), which is a necessary and sufficient condition to obtain detailed balance (see Appendix E.2 for details about the necessity). As corollary of Theorem 1, we obtain the link between the neutral model with density-dependent immigration (14) and assumption (8) that leads to Pólya splitting distributions.

Corollary 1 *Any Polya splitting distribution (with sum distribution support of the form* $\{0, \ldots, m\}$ *where* $m \in \mathbb{N} \cup \{\infty\}$*) can be obtained as a stationnary distribution of a jumping process with rates* (14) *verifying detailed balance condition* (13)*. Reciprocally, let a jumping process with rates* (14) *with detail balance condition* (13)*. If at least* $\tilde{K}$ *or* $\tilde{I}$ *is null then the stationary distribution is a Polya splitting distribution.*

Proof: Let $\mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}$ be a Pólya splitting distribution with sum distribution support of the form $\{0, \dots, m\}$ where $m \in \mathbb{N} \cup \{\infty\}$. According to Theorem 1 this is the stationary distribution of a jumping process satisfying assumption (9) on $q_j(\boldsymbol{n})$. Let $\pi_j = \theta_j/|\boldsymbol{\theta}|$ and $m(n)$, $d(n)$ such that

$$\frac{m(n)}{d(n+1)} = |\boldsymbol{\theta}| s(n). \tag{15}$$

Moreover, let assume that

$$\tilde{K} = \begin{cases} |\boldsymbol{\theta}|^{-1} & \text{if } c = -1, \\ 0 & \text{if } c = 0 \text{ or } c = 1, \end{cases} \quad \text{and} \quad \tilde{I} = \begin{cases} 0 & \text{if } c = -1 \text{ or } c = 0, \\ |\boldsymbol{\theta}|^{-1} & \text{if } c = 1. \end{cases}$$

Then

$$q_j(\boldsymbol{n}) = \frac{m(|\boldsymbol{n}|)}{d(|\boldsymbol{n}|+1)} \frac{\pi_j + (\tilde{I} - \tilde{K}) n_j}{n_j + 1} \mathbb{1}_{\theta_j + c n_j \geq 0},$$
$$q_j(\boldsymbol{n}) = \frac{m(|\boldsymbol{n}|)\left(\pi_j - \tilde{K} n_j\right) + b(|\boldsymbol{n}|) n_j}{d(|\boldsymbol{n}|+1)(n_j+1)} \mathbb{1}_{\theta_j + c n_j \geq 0},$$

where $b(n) = \tilde{I} m(n)$ and $\mathbb{1}_{\theta_j + c n_j \geq 0} = \mathbb{1}_{\pi_j - \tilde{K} n_j \geq 0}$ and thus the desired result. Reciprocally, let a jumping process with rates (14) with detail balance condition (13). If at least $\tilde{K}$ or $\tilde{I}$ is null then the parametrization is reversible.

In the previous section, we derived results about the variation of $s(n)$ among the nine Pólya splitting distributions presented in table 1. Migration $m(n)$ and death $d(n)$ rates directly relate to $s(n)$ through $s(n) = |\boldsymbol{\theta}|^{-1} \frac{m(n)}{d(n+1)}$. Therefore, when assuming that one of these rates is constant with community size n, one obtains biologically interpretable results about the density-dependence effect on the other rate. For instance, if one assumes that $d(n) = 1$ (i.e., constant per-capita local death or emigration rate), $m(n) = |\boldsymbol{\theta}| s(n)$. Here again discarding the canonical case with $c = 1$ and $\gamma < 1$, we obtain that $m(n)$ always decreases in canonical cases, is constant in independent cases and increases in dependent non-canonical cases. From a biological perspective, this suggests that, as community size increases and before hitting a potential regulation threshold $m < \infty$, immigration and reproduction of within the community become harder in the canonical case (negative density dependence), remain unaffected in the independent case and become easier in the other case (positive density dependence). We also obtain that $m(n)$ is convex when $c = -1$, which suggest that the marginal variation in reproduction and immigration increases as community density increases. Migration $m(n)$ is linear when $c = 0$ and is concave when $c = 1$, which means that the marginal variation in reproduction and immigration increases as community density decreases (see Table 3).

5 Discussion and perspectives

We presented the Polya-splitting distributions, a set of multivariate distributions with two key properties : the sum is a positive random variable on $\mathbb{N}$ and the split conditionally to the sum is a Polya distribution. We recalled that Polya-distribution (the split) can be classified

	$c=-1$	$c=0$	$c=1$
All cases	$\tilde{K}=\lvert\boldsymbol{\theta}\rvert^{-1}$ $\tilde{I}=0$ $d(\lvert\boldsymbol{n}\rvert)=1$	$\tilde{K}=0$ $\tilde{I}=0$ $d(\lvert\boldsymbol{n}\rvert)=1$	$\tilde{K}=0$ $\tilde{I}=\lvert\boldsymbol{\theta}\rvert^{-1}$ $d(\lvert\boldsymbol{n}\rvert)=1$
Canonical cases	$m(\lvert\boldsymbol{n}\rvert)=\lvert\boldsymbol{\theta}\rvert\frac{(m-\lvert\boldsymbol{n}\rvert)}{\gamma+1-(m-\lvert\boldsymbol{n}\rvert)}\mathbb{1}_{\lvert\boldsymbol{n}\rvert\leq m}$ $b(\lvert\boldsymbol{n}\rvert)=0$	$m(\lvert\boldsymbol{n}\rvert)=\frac{\lvert\boldsymbol{\theta}\rvert}{\gamma}(m-\lvert\boldsymbol{n}\rvert)\mathbb{1}_{\lvert\boldsymbol{n}\rvert\leq m}$ $b(\lvert\boldsymbol{n}\rvert)=0$	$m(\lvert\boldsymbol{n}\rvert)=\lvert\boldsymbol{\theta}\rvert\frac{m-\lvert\boldsymbol{n}\rvert}{\gamma-1+m-\lvert\boldsymbol{n}\rvert}\mathbb{1}_{\lvert\boldsymbol{n}\rvert\leq m}$ $b(\lvert\boldsymbol{n}\rvert)=\frac{m-\lvert\boldsymbol{n}\rvert}{\gamma-1+m-\lvert\boldsymbol{n}\rvert}\mathbb{1}_{\lvert\boldsymbol{n}\rvert\leq m}$
Independent cases	$m(\lvert\boldsymbol{n}\rvert)=\alpha\lvert\boldsymbol{\theta}\rvert$ $b(\lvert\boldsymbol{n}\rvert)=0$	$m(\lvert\boldsymbol{n}\rvert)=\alpha\lvert\boldsymbol{\theta}\rvert$ $b(\lvert\boldsymbol{n}\rvert)=0$	$m(\lvert\boldsymbol{n}\rvert)=\alpha\lvert\boldsymbol{\theta}\rvert$ $b(\lvert\boldsymbol{n}\rvert)=\alpha$
Dependent non-canonical cases	$m(\lvert\boldsymbol{n}\rvert)=\lvert\boldsymbol{\theta}\rvert\frac{a+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+b-\lvert\boldsymbol{n}\rvert-1}$ $b(\lvert\boldsymbol{n}\rvert)=0$	$m(\lvert\boldsymbol{n}\rvert)=\lvert\boldsymbol{\theta}\rvert\frac{a+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+b}$ $b(\lvert\boldsymbol{n}\rvert)=0$	$m(\lvert\boldsymbol{n}\rvert)=\lvert\boldsymbol{\theta}\rvert\frac{b+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+a+b+\lvert\boldsymbol{n}\rvert}$ $b(\lvert\boldsymbol{n}\rvert)=\frac{b+\lvert\boldsymbol{n}\rvert}{\lvert\boldsymbol{\theta}\rvert+a+b+\lvert\boldsymbol{n}\rvert}$

Table 3: Parametric hypothesis on extended neutral models leading to the nine Pólya splitting distributions presented in table 1. Dark gray indicates cases where the split distribution has been characterized (Haegeman and Etienne, 2008) while light gray stands for cases where both split and sum distributions have been characterized (Etienne et al., 2007). Other cases (white boxes) are new.

in three categories, depending on a parameter c = -1, 0 or 1. The case c=1 corresponds to a Dirichlet-multinomial split, the case c=0 to a multinomial split and the case c=-1 to a hypergeometric split. Our main contribution is to connect those distributions to the neutral theory of biodiversity in ecology, a useful null model allowing the evaluation of non-neutral processes such as adaptation or natural selection (Alonso et al., 2006). We found that for any Polya-splitting distribution, irrespective of the value of c in the split but with a sum support of the form $\{0,\ldots,m\}$ where $m\in\mathbb{N}\cup\{\infty\}$, there exists a multivariate jump process of neutral species with such stationnary distribution. However, staying at the very general level for the sum distribution, the associated transition rates may not have a straightforward biological interpretation. We therefore exhibited, nine transition rates parametrization with meaningful biological interpretation leading to usual parametric distributions. Reciprocally, if multivariate jump process of neutral species follows the detailed balance assumption and has a well-defined stationary distribution of the sum, then the multivariate distribution of species counts is a Polya-splitting distribution.

Kadmon and Allouche (2007) had already shown that a variety of neutral jump processes with detailed balance assumption (formalized in eq. 2 in their work) could generate Multinomial or Dirichlet-Multinomial split distributions that correspond to Polya distributions with $c=0$ or 1 respectively. However, they did not identify neutral jump processes that could generate $c=-1$ because they made the classic assumption that the positive jump rate of a species j linearly increased with the local population size of species j with a positive or null slope b that corresponds to a per-capita bitrh rate. Under this assumption, only Multinomial and Dirichlet-multinomial split can be obtained ($c=0$ or 1). Because we started from the description of the full family of Polya-distribution, we were able to ask the question of whether Polya-splitting distribution with hypergeometric split ($c=-1$) could also be obtained as stationnary distributions of neutral jump processes — which we showed to be true — and what were the peculiarities of these processes that depart from e.g. models condidered in Kadmon and Allouche (2007). We evidenced that the key difference is that hypergeometric split when the positive jump rate of species j linearly decreases with its local population size. This would correspond to a negative per capita birth rate in Kadmon and

Allouche (2007), hence explaining why they did not explore this track. Here, we showed that this negative relationship between positive jump rate and local population size can emerge in a well-defined case: when a limited quantity of available propagules can immigrate from the regional pool to a local site where local birth rate is zero, *i.e.* the community is a pure sink.

We focused our study on the stationary state of a biological community following a neutral jump process. In practice, only a fraction of the community is observed, through a sampling process. In the case $c = 0$ or 1, Etienne and Alonso (2005) noticed that the Dirichlet-multinomial split of the sum among species verify a "subsampling property": when applying a hypergeometric sampling process with fixed sample size over the community, the resulting subsample still followed a Dirichlet-multinomial distribution with the same parameters. We conjecture that this property still holds for the case $c = -1$, although this remains to be properly shown. This property remarkably simplifies the statistical study of species relative abundances within the community. By contrast, the hypergeometric sampling with fixed size does not allow studying the sum distribution, because the sample size is artificially controlled, independently from the real community size. This is quite limiting: our results emphasized the importance of studying the total sum abundance as a random variable, because of its links with the dependence structures among species. In particular, we show that independence is a consequence of parametric assumption made on birth and death rates and not a necessary assumption *per se*, contrary to what was posited by other authors (Etienne et al., 2007). Therefore we suggest that a stronger focus should now be given on sampling processes that preserves information about the community size, like process that controls the distance covered or the time spent during sampling rather than the number of individuals. This requires to explicit sampling models accounting for the spatio-temporal distribution of studied organisms (Jousimo and Ovaskainen, 2016) and to study stability properties of associated thinning operators (Peyhardi, 2023).

Pólya splitting distributions induce only two types of dependence structures: either all species are independent or fully dependent with homogenized correlation sign. To extend this binary setting towards more complex nested dependence structures between species or communities, we suggest the use of recursive application of splitting distributions. From this perspective, the strong closure under addition property plays an essential role by preserving distributions across levels, hence allowing a full control of generated dependencies. This emphasizes the importance of using appropriate choices of Pólya splitting distributions at each level to ensure strong closure under addition.

For instance, let us assume that we aim at simulating species communities composed of five species $s_1, \ldots, s_5$. The first three species (s_1, s_2 and s_3) belongs to a community C_1 and are all fully positively correlated, corresponding to mutualist species context or species sharing underlying environmental factors. The two others species (s_4 and s_5) belongs to a second community, C_2 correlated to C_1. These two species s_4 and s_5 are assumed to be independent. Such situation can be easily obtained combining negative binomial for the sum distributions and multinomial or Dirichlet multinomial distributions for split components (with specific constraint on parameters); see figure 1). Another example relies on the simulation of negatively correlated species within the first community C_1 (exclusive species context) and non-dependent species within C_2 and assuming independence between C_1 and C_2. Such simulation can be performed combining binomial distributions for sum parts with

multinomial and/or hypergeometric distribution for splits. However, does such sequential structure is solution of the master equation? It remains an open question and should be carefully study.

While multinomial or Dirichlet multinomial regression has already been used in neutral community analysis (Jabot et al., 2008; Jabot, 2010), the inclusion of environmental factors in the Pólya splitting distributions is a natural extension. It could be achieved assuming parameters varied according to covariates as follows:

$$\mathcal{P}_{\Delta_n}^{[c]}\{\boldsymbol{\theta}(\boldsymbol{x})\} \underset{n}{\wedge} \mathcal{L}\{\psi(\boldsymbol{x})\},$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)$ denotes the vector explanatory variables (see Peyhardi et al. (2021) for more details in the multinomial splitting regression context). Note that parameters could be constrained to be the same in the split and sum parts.

Finally, combining graph hierarchical approach with the inclusion of environmental covariates at each node leads to propose nested multi-level inhomogenous splitting models. Such models should be interesting alternatives to classical approaches used in joint species distribution contexts mainly based on conditional Independence's (Warton et al., 2015; Ovaskainen and Abrego, 2020) and the use of the multivariate Poisson log-Normal distribution. Comparatively, our approach allows to model dependencies between species at the observation level, while the classical JSDM's model dependencies at the latent process strata. While correlation relationships estimated at the latent processes inform correlations between observations, it does not allow to deduce dependencies structures at the observation scales (Aïtchison and Ho, 1989; Chiquet et al., 2021). A null correlation does not imply independence in the multivariate Poisson context.

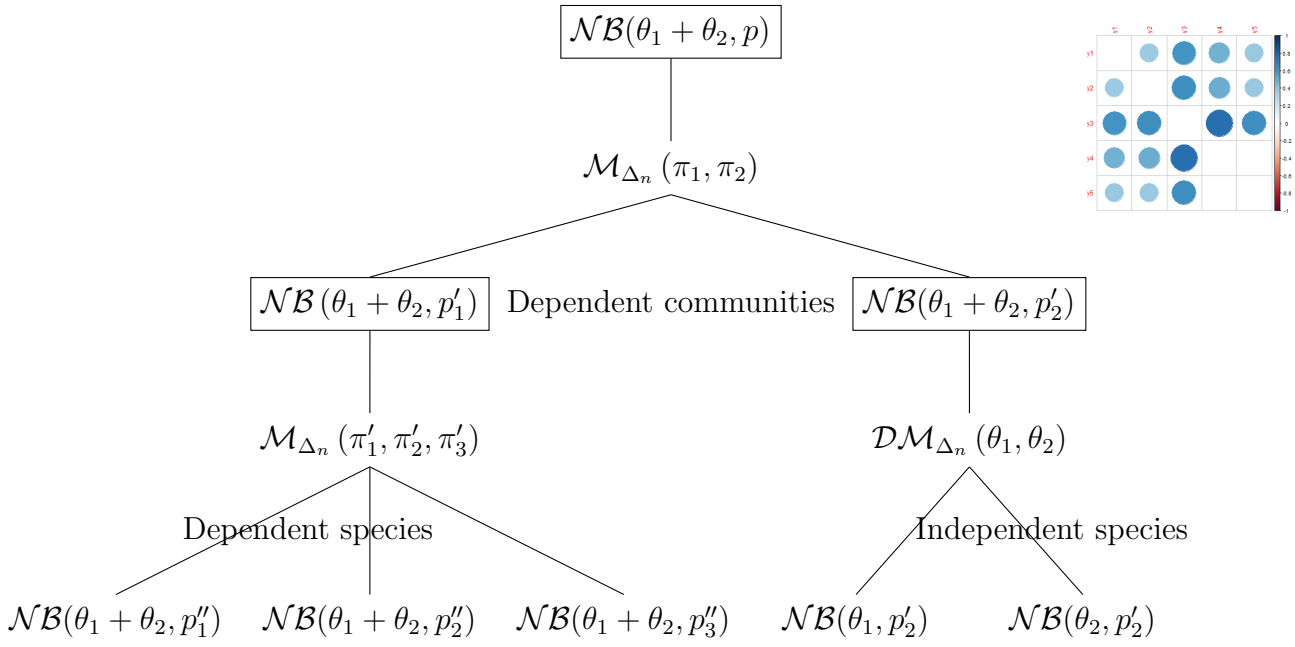

Figure 1: Example of possible simulation schemes combining multi dependent levels

Contribution

F. Laroche: Conceptualization, Formal analysis, Writing - Original Draft **F. Mortier**: Conceptualization, Writing - Original Draft, Funding acquisition. **J Peyhardi**: Methodology, Formal analysis, Writing - Original Draft.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the GAMBAS project funded by the French National Research Agency (ANR-18-CE02-0025). F. Laroche was supported by the BloBiForM project funded by French National Research Agency (ANR-19-CE32-0002-01).

Declarations

References

Aïtchison J, Ho C (1989) The multivariate poisson-log normal distribution. Biometrika 76:643–653

Alonso D, Etienne RS, McKane AJ (2006) The merits of neutral theory. Trends in Ecology & Evolution 21(8):451–457, DOI 10.1016/j.tree.2006.03.019, URL `http://www.sciencedirect.com/science/article/pii/S0169534706001650`

Bell G (2005) The co-distribution of species in relation to the neutral theory of community ecology. Ecology 86:1757–1770

Canard E, Mouquet N, Marescot L, Gaston K, Gravel D, Mouillot D (2012) Emergence of structural patterns in neutral trophic networks. Plos One 7(8), URL `https://doi.org/10.1371/journal.pone.0038295`

Chiquet J, Mariadassou M, Robin S (2021) The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. Frontiers in Ecology and Evolution

9:188, DOI 10.3389/fevo.2021.588292, URL https://www.frontiersin.org/article/10.3389/fevo.2021.588292

Donnelly P, Nordborg M, Joyce P (2001) Likelihoods and Simulation Methods for a Class of Nonneutral Population Genetics Models. Genetics 159(2):853, URL http://www.genetics.org/content/159/2/853.abstract

Etienne R, Alonso D (2005) A dispersal-limited sampling theory for species and alleles. Ecology Letters 8(11):1147–1156, DOI 10.1111/j.1461-0248.2005.00817.x, wOS:000232535300003

Etienne R, Olff H (2004) A novel genealogical approach to neutral biodiversity theory. Ecology Letters 7(3):170–175, DOI 10.1111/j.1461-0248.2004.00572.x, wOS:000189232400002

Etienne R, Alonso D, McKane A (2007) The zero-sum assumption in neutral biodiversity theory. Journal of theoretical biology 248(3):522–536

Haegeman B, Etienne R (2008) Relaxing the zero-sum assumption in neutral biodiversity theory. Journal of Theoretical Biology 252(2):288–294

Harris K, Parsons T, Ijaz U, Lahti L, Holmes I, Quince C (2017) Linking Statistical and Ecological Theory: Hubbell's Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process. Proceedings of the IEEE 105(3):516–529, DOI 10.1109/JPROC.2015.2428213

Hubbell S (2001) The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, NJ

Irwin J (1968) The generalized Waring distribution applied to accident theory. Journal of the Royal Statistical Society Series A (General) pp 205–225

Jabot F (2010) A stochastic dispersal-limited trait-based model of community dynamics. Journal of Theoretical Biology 262(4):650–661, DOI https://doi.org/10.1016/j.jtbi.2009.11.004, URL https://www.sciencedirect.com/science/article/pii/S002251930900530X

Jabot F, Etienne RS, Chave J (2008) Reconciling neutral community models and environmental filtering: theory and an empirical test. Oikos 117(9):1308–1320, DOI https://doi.org/10.1111/j.0030-1299.2008.16724.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0030-1299.2008.16724.x, https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0030-1299.2008.16724.x

Jousimo J, Ovaskainen O (2016) A spatio-temporally explicit random encounter model for large-scale population surveys. PloS one 11:e0162447, DOI 10.1371/journal.pone.0162447

Kadmon R, Allouche O (2007) Integrating the effects of area, isolation, and habitat heterogeneity on species diversity: A unification of island biogeography and niche theory. The American naturalist 170:443–54, DOI 10.1086/519853

Laroche F, Jarne P, Lamy T, David P, FA M (2015) A neutral theory for interpreting correlations between species and genetic diversity in communities. The American Naturalist 185(1):59–69

Laroche F, Violle C, Taudière A, Munoz F (2020) Analyzing snapshot diversity patterns with the neutral theory can show functional groups' effects on community assembly. Ecology 101(4):e02977, DOI https://doi.org/10.1002/ecy.2977, URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.2977, https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecy.2977

Ovaskainen O, Abrego N (2020) Joint Species Distribution Modelling: With Applications in R. Ecology, Biodiversity and Conservation, Cambridge University Press, DOI 10.1017/9781108591720

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, Fernique P, Durand JB (2021) Splitting models for multivariate count data. Journal of Multivariate Analysis 181:104677

Warton D, Blanchet F, O'Hara R, Ovaskainen O, Taskinen S, Walker S, Hui F (2015) So many variables: Joint modeling in community ecology. Trends in Ecology and Evolution 30:766–779

Xekalaki E (1986) The multivariate generalized Waring distribution. Communications in Statistics - Theory and Methods 15(3):1047–1064

A Notations of different distributions

Name	Notation	space parameter	support	pmf (p_n)
Pólya	$\mathcal{P}_m^{[c]}(\theta,\gamma)$	$\theta\in\Theta_c$, $\gamma\in\Theta_c$, $m\in\mathbb{N}^*$		$\dfrac{R_\theta^{[c]}(n)R_\gamma^{[c]}(m-n)}{R_{\theta+\gamma}^{[c]}(m)}$
hypergeometric $(c=-1)$	$\mathcal{H}_m(k,l)$	$k\in\mathbb{N}^*$, $l\in\mathbb{N}^*$,$m\in\mathbb{N}^*$, $m\leq k+l$	$\{0,\dots,m\}\cap\{m-l,\dots,k\}$	$\dfrac{\binom{k}{n}\binom{l}{m-n}}{\binom{k+l}{m}}$
binomial $(c=0)$	$\mathcal{B}_m(p)$	$p\in(0,1)$, $m\in\mathbb{N}^*$	$\{0,\dots,m\}$	$\binom{m}{n}p^n(1-p)^{m-n}$
beta binomial $(c=1)$	$\beta\mathcal{B}_m(a,b)$	$a\in\mathbb{R}_+^*$, $b\in\mathbb{R}_+^*$,$m\in\mathbb{N}^*$	$\{0,\dots,m\}$	$\dfrac{\binom{n+a-1}{n}\binom{m-n+b-1}{m-n}}{\binom{m+a+b-1}{m}}$

Table A.1: Notations and pmf of univariate Pólya distributions

Name	Notation	space parameter	support	pmf $(p_{\boldsymbol{n}})$
Singular version				
multivariate Pólya	$\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta})$	$n\in\mathbb{N}^*$, $\boldsymbol{\theta}\in\Theta_c^J$		$\dfrac{\prod_{j=1}^J R_{\theta_j}^{[c]}(n_j)}{R_{\lvert\boldsymbol{\theta}\rvert}^{[c]}(n)}$
multivariate hypergeometric $(c=-1)$	$\mathcal{H}_{\Delta_n}(\boldsymbol{k})$	$n\in\mathbb{N}^*$, $\boldsymbol{k}\in\mathbb{N}^{*J}$	$\Delta_n\cap\blacksquare_{\boldsymbol{k}}$	$\dfrac{\prod_{j=1}^J\binom{k_j}{n_j}}{\binom{\lvert\boldsymbol{k}\rvert}{n}}$
multinomial $(c=0)$	$\mathcal{M}_{\Delta_n}(\boldsymbol{\pi})$	$n\in\mathbb{N}^*$, $\boldsymbol{\pi}\in\Delta$	Δ_n	$\binom{n}{\boldsymbol{n}}\prod_{j=1}^J\pi_j^{n_j}$
Dirichlet multinomial $(c=1)$	$\mathcal{DM}_{\Delta_n}(\boldsymbol{\alpha})$	$n\in\mathbb{N}^*$, $\boldsymbol{\alpha}\in\mathbb{R}_+^{*J}$	Δ_n	$\dfrac{\prod_{j=1}^J\binom{n_j+\alpha_j-1}{n_j}}{\binom{n+\lvert\boldsymbol{\alpha}\rvert-1}{n}}$
Non-singular version				
multivariate Pólya	$\mathcal{P}_{\blacktriangle_m}^{[c]}(\boldsymbol{\theta},\gamma)$	$m\in\mathbb{N}^*$, $\boldsymbol{\theta}\in\Theta_c^J$, $\gamma\in\Theta_c$		$\dfrac{R_\gamma^{[c]}(m-\lvert\boldsymbol{n}\rvert)\prod_{j=1}^J R_{\theta_j}^{[c]}(n_j)}{R_{\lvert\boldsymbol{\theta}\rvert+\gamma}^{[c]}(m)}$
multivariate hypergeometric $(c=-1)$	$\mathcal{H}_{\blacktriangle_m}(\boldsymbol{k},l)$	$m\in\mathbb{N}^*$, $\boldsymbol{k}\in\mathbb{N}^{*J}$, $l\in\mathbb{N}^*$ $m\leq\lvert\boldsymbol{k}\rvert+l$	$(\blacktriangle_m\setminus\blacktriangle_{m-l})\cap\blacksquare_{\boldsymbol{k}}$	$\dfrac{\binom{l}{m-\lvert\boldsymbol{n}\rvert}\prod_{j=1}^J\binom{k_j}{n_j}}{\binom{\lvert\boldsymbol{k}\rvert+l}{m}}$
multinomial $(c=0)$	$\mathcal{M}_{\blacktriangle_m}(\boldsymbol{\pi}^*)$	$m\in\mathbb{N}^*$, $\boldsymbol{\pi}^*\in\blacktriangle$	$\blacktriangle_m$	$\binom{m}{\boldsymbol{n}}(1-\lvert\boldsymbol{\pi}^*\rvert)^{m-\lvert\boldsymbol{n}\rvert}\prod_{j=1}^J\pi_j^{n_j}$
Dirichlet multinomial $(c=1)$	$\mathcal{DM}_{\blacktriangle_m}(\boldsymbol{\alpha},\beta)$	$m\in\mathbb{N}^*$, $\boldsymbol{\alpha}\in\mathbb{R}_+^{*J}$, $\beta\in\mathbb{R}_+^*$	$\blacktriangle_m$	$\dfrac{\binom{m-\lvert\boldsymbol{n}\rvert+\beta-1}{m-\lvert\boldsymbol{n}\rvert}\prod_{j=1}^J\binom{n_j+\alpha_j-1}{n_j}}{\binom{m+\lvert\boldsymbol{\alpha}\rvert+\beta-1}{m}}$

Table A.2: Notations and pmf of multivariate Pólya distributions (singular and non-singular versions)

Name	Notation	space parameter	support	pmf (p_n)
Poisson	$\mathcal{P}(\lambda)$	$\lambda \in \mathbb{R}_+^*$	$\mathbb{N}$	$e^{-\lambda}\dfrac{\lambda^n}{n!}$
negative binomial	$\mathcal{NB}(r,p)$	$r \in \mathbb{R}_+^*$, $p \in (0,1)$	$\mathbb{N}$	$\binom{n+r-1}{n} p^n (1-p)^r$
beta-negative binomial	$\beta\mathcal{NB}(r,a,b)$	$r \in \mathbb{R}_+^*$, $a \in \mathbb{R}_+^*$, $b \in \mathbb{R}_+^*$	$\mathbb{N}$	$\binom{n+r-1}{n} \dfrac{B(a+r,b+n)}{B(a,b)}$

Table A.3: Notations and pmf of some usual univariate distributions

B Strong closure under addition (specific cases)

Let us show the strong closure under addition for the three Pólya splitting distributions presented in the third line of Table 1. We have to show that marginal distributions and sum distribution belong to the same family.

- $\underline{c=-1}$ According to Theorem 1 of Peyhardi et al. (2021) the marginals are given by the hypergeometric damage distribution

$$\begin{aligned}
\mathcal{H}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \beta\mathcal{B}_{|\boldsymbol{\theta}|}(a,b) &= \mathcal{H}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \left\{ \mathcal{B}_{|\boldsymbol{\theta}|}(p) \underset{p}{\wedge} \beta(a,b) \right\} \\
&= \left\{ \mathcal{H}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \mathcal{B}_{|\boldsymbol{\theta}|}(p) \right\} \underset{p}{\wedge} \beta(a,b) \\
&= \mathcal{B}_{\theta_j}(p) \underset{p}{\wedge} \beta(a,b) \\
\mathcal{H}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \beta\mathcal{B}_{|\boldsymbol{\theta}|}(a,b) &= \beta\mathcal{B}_{\theta_j}(a,b)
\end{aligned}$$

  The first equality uses the definition of the beta-binomial distribution, the second one uses the Fubini theorem (inversion of sum on n and integral on p), the third one uses the stability of the binomial distribution under hypergeometric damage process (obtained in the case of independence) and the last one uses again the definition of the beta-binomial distribution.

- $\underline{c=0}$ According to Theorem 1 of Peyhardi et al. (2021) the marginals are given by the binomial damage distribution $\mathcal{B}_n(\pi_j) \underset{n}{\wedge} \mathcal{NB}(r,p)$. Theorem 6 of Peyhardi et al. (2021) showed the stability of the negative binomial distribution under the binomial damage process, i.e., we have

$$\mathcal{B}_n(\pi_j) \underset{n}{\wedge} \mathcal{NB}(r,p) = \mathcal{NB}(r,p'),$$

  where $p' = \frac{\pi_j p}{\pi_j p + 1 - p}$. The demonstration is based on the generative function of a binomial damage distribution. This result can also be obtained by following the way of the demonstartion of the previous case, recalling that a negative binomial is a Poisson mixed by a gamma distribution.

- $\underline{c=1}$ According to Theorem 1 of Peyhardi et al. (2021) the marginals are given by the beta-binomial damage distribution

$$\begin{aligned}
\beta\mathcal{B}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \beta\mathcal{NB}(|\boldsymbol{\theta}|, a, b) &= \beta\mathcal{B}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \left\{ \mathcal{NB}(|\boldsymbol{\theta}|, p) \underset{p}{\wedge} \beta(a, b) \right\} \\
&= \left\{ \beta\mathcal{B}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \mathcal{NB}(|\boldsymbol{\theta}|, p) \right\} \underset{p}{\wedge} \beta(a, b) \\
&= \mathcal{NB}(\theta_j, p) \underset{p}{\wedge} \beta(a, b) \\
\beta\mathcal{B}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \beta\mathcal{NB}(|\boldsymbol{\theta}|, a, b) &= \beta\mathcal{NB}(\theta_j, a, b)
\end{aligned}$$

  This demonstration follows the same ways as in case $c = -1$.

C Some stationary distributions of univariate birth-death process

According to different assumptions on the ratio $q(n) = \frac{q^+(n)}{q^-(n+1)}$, we find different distributions $(p_n)_{n\geq 0}$. Recall that

$$p_n = \frac{Q_n}{\sum_{m\geq 0} Q_m},$$

with $Q_n = \prod_{k=0}^{n-1} q(k)$.

C.1 Univariate Pólya distributions

- $\underline{\text{Hypergeometric distribution:}}$ if $q(n) = \frac{(k-n)(m-n)}{(n+1)(l-m+n+1)} \mathbf{1}_{\max(0,m-l)\leq n<\min(m,k)}$ with $k \in \mathbb{N}^*, l \in \mathbb{N}^*$ and $m \leq k + l$ then $Q_n = \frac{\binom{k}{n}\binom{l}{m-n}}{\binom{l}{m}}$ and

$$p_n = \frac{\binom{k}{n}\binom{l}{m-n}}{\binom{k+l}{m}},\ \max(0, m-l) \leq n \leq \min(m, k),$$

  i.e., $N \sim \mathcal{H}_m(k, l)$.

- $\underline{\text{Binomial distribution:}}$ if $q(n) = \frac{(M-n)\pi}{(n+1)(1-\pi)} \mathbb{1}_{n<m}$ with $\pi \in (0, 1)$ and $m \in \mathbb{N}^*$ then $Q_n = \binom{m}{n} \frac{\pi^n}{(1-\pi)^n}$ and

$$p_n = \binom{m}{n} \pi^n (1-\pi)^{m-n},\ 0 \leq n \leq m$$

  i.e., $N \sim \mathcal{B}_m(\pi)$.

- Beta-binomial distribution: if $q(n) = \frac{(a+n)(m-n)}{(n+1)(m+b-n-1)}\mathbb{1}_{n<m}$ with $a \in \mathbb{R}_+^*, b \in \mathbb{R}_+^*$ and $m \in \mathbb{N}^*$ then $Q_n = \frac{\binom{n+a-1}{n}\binom{m-n+b-1}{m-n}}{\binom{m+b-1}{m}}$ and

$$p_n = \frac{\binom{n+a-1}{n}\binom{m-n+b-1}{m-n}}{\binom{m+a+b-1}{m}},\ n \leq m$$

  i.e., $N \sim \beta\mathcal{B}_m(a, b)$.

C.2 Other univariate distributions

- Poisson distribution: if $q(n) = \frac{1}{n+1}\alpha$ with $\alpha \in \mathbb{R}_+^*$ then $Q_n = \frac{\alpha^n}{n!}$ and

$$p_n = e^{-\alpha}\frac{\alpha^n}{n!},\ n \geq 0,$$

  i.e., $N \sim \mathcal{P}(\alpha)$.

- Negative binomial distribution: if $q(n) = \frac{n+\alpha}{n+1}\pi$ with $\pi \in (0, 1)$ and $\alpha \in (0, \infty)$ then $Q_n = \binom{n+\alpha-1}{n}\pi^n$ and

$$p_n = \binom{n+\alpha-1}{n}\pi^n(1-\pi)^\alpha,\ n \geq 0,$$

  i.e., $N \sim \mathcal{NB}(\alpha, \pi)$.

  Geometric distribution: if $q(n) = \pi$ with $\pi \in (0, 1)$ then $Q_n = \pi^n$ and

$$p_n = \pi^n(1-\pi),\ n \geq 0,$$

  i.e., $N \sim \mathcal{G}(\pi)$.

  Remark that the geometric distribution is a sub-case of the negative-binomial distribution, more precisely we have $\mathcal{G}(\pi) = \mathcal{NB}(1, \pi)$.

- Beta negative binomial distribution: if $q(n) = \frac{(\alpha+n)(b+n)}{(n+1)(\alpha+a+b+n)}$ with $\alpha \in \mathbb{R}_+^*, b \in \mathbb{R}_+^*$ and $a \in \mathbb{R}_+^*$ then $Q_n = \binom{n+\alpha-1}{n}\frac{\Gamma(b+n)\Gamma(\alpha+a+b)}{\Gamma(b)\Gamma(\alpha+a+b+n)} = \binom{n+\alpha-1}{n}\frac{B(a+\alpha,b+n)}{B(\alpha+a,b)}$ and

$$p_n = \binom{n+\alpha-1}{n}\frac{B(a+\alpha, b+n)}{B(a, b)},$$

  where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, i.e., $N \sim \beta\mathcal{NB}(\alpha, a, b)$.

  Remark that if the parameters a and b are positive integers, then the beta-binomial distribution turns out to be the negative hypergeometric distribution. Otherwise, the beta negative binomial distribution is also called the generalized waring distribution (Irwin, 1968).

D Parametric hypothesis on $s(n)$ for the canonical case

Assume that $s(n) = 1/r_\gamma^{[c]}(m-n-1)$ for some $\gamma \in \Theta$ and $m \in \mathbb{N}$. Then we have

$$\begin{aligned}
\prod_{k=0}^{n-1} s(k) &= \frac{1}{\prod_{k=0}^{n-1} r_\gamma^{[c]}(m-k-1)} \\
&= \frac{1}{r_\gamma^{[c]}(m-1) \times \cdots \times r_\gamma^{[c]}(m-n)} \\
&= \frac{r_\gamma^{[c]}(m-n-1) \times \cdots \times r_\gamma^{[c]}(0)}{r_\gamma^{[c]}(m-1) \times \cdots \times r_\gamma^{[c]}(0)} \\
\prod_{k=0}^{n-1} s(k) &= \frac{R_\gamma^{[c]}(m-n)}{R_\gamma^{[c]}(m)}
\end{aligned}$$

Therefore

$$\begin{aligned}
\sum_{n=0}^{m} R_{|\boldsymbol{\theta}|}^{[c]}(n) \prod_{k=0}^{n-1} s(k) &= \frac{1}{R_\gamma^{[c]}(m)} \sum_{n=0}^{m} R_{|\boldsymbol{\theta}|}^{[c]}(n) R_\gamma^{[c]}(m-n) \\
&= \frac{1}{R_\gamma^{[c]}(m)} (R_{|\boldsymbol{\theta}|}^{[c]} * R_\gamma^{[c]})(m) \\
\sum_{n=0}^{m} R_{|\boldsymbol{\theta}|}^{[c]}(n) \prod_{k=0}^{n-1} s(k) &= \frac{1}{R_\gamma^{[c]}(m)} R_{|\boldsymbol{\theta}|+\gamma}^{[c]}(m)
\end{aligned}$$

Finally the pmf of the sum given by (11) becomes

$$P(|\boldsymbol{N}| = n) = \frac{R_{|\boldsymbol{\theta}|}^{[c]}(n) R_\gamma^{[c]}(m-n)}{R_{|\boldsymbol{\theta}|+\gamma}^{[c]}(m)},$$

for all $n \leq m$ and zero otherwise. This is the pmf of the univariate Pólya distribution $\mathcal{P}_m^{[c]}(|\boldsymbol{\theta}|, \gamma)$. Therefore the multivariate stationary distribution is the non-singular version of the Pólya distribution thanks to the identity

$$\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{P}_m^{[c]}(|\boldsymbol{\theta}|, \gamma) = \mathcal{P}_{\blacktriangle_m}^{[c]}(\boldsymbol{\theta}, \gamma).$$

According to Theorem 1 of Peyhardi et al. (2021) the marginals are given by the damage distributions

$$\mathcal{P}_n^{[c]}(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \mathcal{P}_m^{[c]}(|\boldsymbol{\theta}|, \gamma) = \mathcal{P}_m^{[c]}(\theta_j, |\boldsymbol{\theta}| + \gamma),$$

for all $j = 1, \ldots, J$.

E Necessary and sufficient condition for detailed balance in the classic neutral theory and with density-dependent migration

E.1 Classic neutral theory

Recall the assumptions of classic neutral theory regarding the jumping rates (12):

$$\begin{aligned} q_j^+(\boldsymbol{n}) &= m(|\boldsymbol{n}|)\pi_j + n_j b(|\boldsymbol{n}|) \\ q_j^-(\boldsymbol{n}) &= n_j d(|\boldsymbol{n}|) \end{aligned}$$

where we assume that $n > 0 \implies d(n) > 0$ (i.e. no individual is immortal).

We assume that the above process has a stationnary distribution with support $\mathbb{S} \subset \mathbb{N}^J$. Because no individual is immortal, $\mathbf{0} \in \mathbb{S}$ and if $\boldsymbol{n} \in \mathbb{S}$ and $\boldsymbol{n'} \in \mathbb{N}^J \,|\, \forall i \in \{1, ..., J\}$, $n'_i \leq n_i$ then $\boldsymbol{n'} \in \mathbb{S}$.

Define $K_m = \min\{n \in \mathbb{N} | m(n) = 0\}$ and $K = \min\{n \in \mathbb{N} | m(n) = b(n) = 0\}$. By definition, $K_m \leq K \leq +\infty$. If $K_m = 0$ then $\mathbb{S} = \{\mathbf{0}\}$. If $K_m > 0$ then $\mathbb{S} = \blacktriangle_K$. In what follows we assume that $K_m > 0$ and $\mathbb{S} = \blacktriangle_K$.

We seek for necessary conditions to obtain detailed balance of the stationary distribution which is depicted by the Kolmogrorov criterion (7):

$$\forall \boldsymbol{n} \in \blacktriangle_K,\ \forall (i,j) \in \{1, ..., J\}^2,\ q_i(\boldsymbol{n})\, q_j(\boldsymbol{n} + \boldsymbol{e_i}) = q_j(\boldsymbol{n})\, q_i(\boldsymbol{n} + \boldsymbol{e_j}),$$

where $q_j(\boldsymbol{n}) = \frac{q_j^+(\boldsymbol{n})}{q_j^-(\boldsymbol{n}+\boldsymbol{e_j})}$. Recall that q_j is well defined because no individual is immortal.

Using the expression of jumping rates, the Kolmogorov criterion becomes :

$$\begin{aligned} &[m(|\boldsymbol{n}|)\pi_i + n_i b(|\boldsymbol{n}|)]\ [m(|\boldsymbol{n}|+1)\pi_j + n_j b(|\boldsymbol{n}|+1)] \\ &= [m(|\boldsymbol{n}|)\pi_j + n_j b(|\boldsymbol{n}|)]\ [m(|\boldsymbol{n}|+1)\pi_i + n_i b(|\boldsymbol{n}|+1)]\,, \end{aligned}$$

which can be simplified as :

$$(\pi_i n_j - \pi_j n_i)\,[m(|\boldsymbol{n}|)b(|\boldsymbol{n}|+1) - m(|\boldsymbol{n}|+1)b(|\boldsymbol{n}|)] = 0$$

which implies in turn that :

$$\forall n \in \{1, ..., K-1\},\ m(n)b(n+1) = m(n+1)b(n) \tag{16}$$

Declining constraint (16) along possible initializations of $m(n)$ and $b(n)$ yields:

- if $m(1) = b(1) = 0$: $K = 1$ and constraint (16) disappears.
- if $m(1) = 0$ and $b(1) > 0$: $\forall n \in \{1, ..., K-1\}, m(n) = 0$
- if $m(1) > 0$ and $b(1) = 0$: $\forall n \in \{1, ..., K-1\}, b(n) = 0$
- if $m(1) > 0$ and $b(1) > 0$: $\forall n \in \{1, ..., K-1\}$: $m(n) > 0, b(n) > 0$ and :

$$b(n) = \tilde{I} m(n)$$

where $\tilde{I} = \frac{b(1)}{m(1)}$.

In summary, we have shown that the stationnary distribution verifies detailed balance only if one of the following conditions holds

- $K_m = 0$
- $K_m = 1$ and $\forall n \in \{1, ..., K-1\}, m(n) = 0$
- $K_m > 1$ and $\exists\, \tilde{I} \geq 0 \,|\, \forall n \in \{1, ..., K_m\}, b(n) = \tilde{I} m(n)$

Reciprocally, it is straightforward to show that each of these conditions is sufficient to obtain detailed balance of the stationary distribution.

E.2 Neutral theory with density-dependent migration

We now turn to the extension of neutral theory including density-dependent immigration as defined in equation (14), which we recall here :

$$\begin{aligned} q_j^+(\boldsymbol{n}) &= \left[m(|\boldsymbol{n}|) \left(\pi_j - \tilde{K} n_j \right) + n_j b(|\boldsymbol{n}|) \right] \mathbb{1}_{\pi_j - \tilde{K} n_j \geq 0} \\ q_j^-(\boldsymbol{n}) &= n_j d(|\boldsymbol{n}|) \end{aligned}$$

with $\tilde{K} \in \mathbb{R}_+$. We define $K_j = \max\{n_j \in \mathbb{N} | \pi_j - \tilde{K} n_j > 0\} + 1$ and $\blacksquare_{\tilde{K}}$ the hypercube $\{0, ..., K_1\} \times \{0, ..., K_2\} \times ... \times \{0, ..., K_J\}$.

We also define K_m and K like in previous section and we assume that $K_m > 0$. Then the support of the stationary distribution is $\mathbb{S} = \blacksquare_{\tilde{K}} \cap \blacktriangle_K$

We seek for necessary conditions to obtain detailed balance of the stationary distribution which is depicted by the Kolmogrorov criterion (7). Using the expression of jumping rates in (14), the Kolmogorov criterion becomes :

$$\begin{aligned} &\left[m(|\boldsymbol{n}|) \left(\pi_i - \tilde{K} n_i \right) + n_i b(|\boldsymbol{n}|) \right] \left[m(|\boldsymbol{n}|+1) \left(\pi_j - \tilde{K} n_j \right) + n_j b(|\boldsymbol{n}|+1) \right] \mathbb{1}_{\pi_i - \tilde{K} n_i \geq 0 \cap \pi_j - \tilde{K} n_j \geq 0} \\ &= \left[m(|\boldsymbol{n}|) \left(\pi_i - \tilde{K} n_i \right) + n_i b(|\boldsymbol{n}|) \right] \left[m(|\boldsymbol{n}|+1) \left(\pi_j - \tilde{K} n_j \right) + n_j b(|\boldsymbol{n}|+1) \right] \mathbb{1}_{\pi_i - \tilde{K} n_i \geq 0 \cap \pi_j - \tilde{K} n_j \geq 0} \end{aligned}$$

which can be simplified as :

$$(\pi_i n_j - \pi_j n_i) \left[m(|\boldsymbol{n}|) b(|\boldsymbol{n}|+1) - m(|\boldsymbol{n}|+1) b(|\boldsymbol{n}|) \right] \mathbb{1}_{\pi_i - \tilde{K} n_i \geq 0 \cap \pi_j - \tilde{K} n_j \geq 0} = 0$$

which implies in turn that :

$$\forall n \in \{0, ..., K' - 1\},\ m(n) b(n+1) = m(n+1) b(n) \tag{17}$$

where $K' = \min\left(\sum_{j=1}^{J} K_j - 1, K \right)$

Declining constraint (17) along possible initializations of $m(n)$ and $b(n)$ yields:

- if $m(1) = b(1) = 0 :\ K = 1$ and constraint (17) disappears.
- if $m(1) = 0$ and $b(1) > 0 : \forall n \in \{1, ..., K' - 1\}, m(n) = 0$
- if $m(1) > 0$ and $b(1) = 0 : \forall n \in \{1, ..., K' - 1\}, b(n) = 0$
- if $m(1) > 0$ and $b(1) > 0 : \forall n \in \{1, ..., K' - 1\} : m(n) > 0, b(n) > 0$ and :

$$b(n) = \tilde{I} m(n)$$

where $\tilde{I} = \frac{b(1)}{m(1)}$. In summary, we have shown that the stationnary distribution verifies detailed balance only if one of the following conditions holds

- $K_m = 0$
- $K_m = 1$ and $\forall n \in \{1, ..., K' - 1\}, m(n) = 0$
- $K_m > 1$ and $\exists\, \tilde{I} \geq 0 \,|\, \forall n \in \{1, ..., K'\}, b(n) = \tilde{I} m(n)$

Reciprocally, it is straightforward to show that each of these conditions is sufficient to obtain detailed balance of the stationary distribution.

F Variation and convexity of $n \to s(n)$

	$c = -1$	$c = 0$	$c = 1$
Canonical cases	$s'(n) = \frac{-(\gamma+1)}{(n-(m-\gamma)+1)^2}\mathbb{1}_{m-\gamma\leq n\leq m}$ $\gamma \in \mathbb{N}^*,\ m \in \mathbb{N}^*,\ m \leq \lvert\boldsymbol{\theta}\rvert + \gamma$	$s'(n) = -\frac{1}{\gamma}\mathbb{1}_{n\leq m}$ $\gamma \in \mathbb{R}_+^*,\ m \in \mathbb{N}^*$	$s'(n) = \frac{-(\gamma-1)}{(\gamma+m-n-1)^2}\mathbb{1}_{n\leq m}$ $\gamma \in \mathbb{R}_+^*,\ m \in \mathbb{N}^*$
	$\implies$ s decreases on $(m-\gamma, m)$	$\implies$ s decreases on $(0, m)$	$\begin{cases}\gamma < 1 \implies s \text{ increases on } (0,m)\\ \gamma > 1 \implies s \text{ decreases on } (0,m)\end{cases}$
	$\implies$ s' increases on $(m-\gamma, m)$	$\implies$ s' constant on $(0, m)$	$\begin{cases}\gamma < \frac{1}{2} \implies s' \text{ increases on } (0, m-1)\\ \frac{1}{2} < \gamma < 1 \implies s' \text{ increases on } (0,m)\\ \gamma > 1 \implies s' \text{ decreases on } (0,m)\end{cases}$
Independent cases	$s'(n) = 0$ $\alpha \in \mathbb{R}_+^*$	$s'(n) = 0$ $\alpha \in \mathbb{R}_+^*$	$s'(n) = 0$ $\alpha \in (0, 1)$
	$\implies$ s constant on $(0, +\infty)$	$\implies$ s constant on $(0, +\infty)$	$\implies$ s constant on $(0, +\infty)$
	$\implies$ s' constant on $(0, +\infty)$	$\implies$ s' constant on $(0, +\infty)$	$\implies$ s' constant on $(0, +\infty)$
Dependent non-canonical cases	$s'(n) = \frac{\lvert\boldsymbol{\theta}\rvert + b + a - 1}{(\lvert\boldsymbol{\theta}\rvert + b - n - 1)^2}\mathbb{1}_{n<\lvert\boldsymbol{\theta}\rvert}$ $a \in \mathbb{R}_+^*,\ b \in \mathbb{R}_+^*$	$s'(n) = \frac{1}{\lvert\boldsymbol{\theta}\rvert + b}$ $a \in \mathbb{R}_+^*$ and $b \in \mathbb{R}_+^*$	$s'(n) = \frac{\lvert\boldsymbol{\theta}\rvert + a}{(\lvert\boldsymbol{\theta}\rvert + a + b + n)^2}$ $a \in \mathbb{R}_+^*$ and $b \in \mathbb{R}_+^*$
	$\implies$ s increases on $(0, \lvert\boldsymbol{\theta}\rvert)$	$\implies$ s constant on $(0, +\infty)$	$\implies$ s increases on $(0, +\infty)$
	$\implies$ s' increases on $(0, \lvert\boldsymbol{\theta}\rvert)$	$\implies$ s' constant on $(0, +\infty)$	$\implies$ s' decreases on $(0, +\infty)$

Table F.1: Variation and convexity of s for the nine Pólya splitting distributions of Table 1