Tree Pólya Splitting distributions for multivariate count data

Samuel Valiquette, Éric Marchand, Jean Peyhardi, Gwladys Toulemonde, Frédéric Mortier

Tree Pólya Splitting distributions for multivariate count data

Samuel Valiquette[1,2,3,4,5] Éric Marchand[3] Jean Peyhardi[4]

Gwladys Toulemonde[4,5] Frédéric Mortier[1,2,6]

[1] UPR Forêts et Sociétés, CIRAD, Montpellier, 34398, France
[2] Forêts et Sociétés, Univ Montpellier, CIRAD, Montpellier, 34398, France
[3] Département de mathématiques, Université de Sherbrooke, Sherbrooke, J1K 2R1, Canada
[4] IMAG, CNRS, Université de Montpellier, Montpellier, 34090, France
[5] LEMON, Inria, Montpellier, 34095, France
[6] Environmental Justice Program, Georgetown University, Washington, D.C., 20007, USA

Abstract

In this article, we develop a new class of multivariate distributions adapted for count data, called Tree Pólya Splitting. This class results from the combination of a univariate distribution and singular multivariate distributions along a fixed partition tree. As we will demonstrate, these distributions are flexible, allowing for the modeling of complex dependencies (positive, negative, or null) at the observation level. Specifically, we present the theoretical properties of Tree Pólya Splitting distributions by focusing primarily on marginal distributions, factorial moments, and dependency structures (covariance and correlations). The abundance of 17 species of Trichoptera recorded at 49 sites is used, on one hand, to illustrate the theoretical properties developed in this article on a concrete

case, and on the other hand, to demonstrate the interest of this type of models, notably by comparing them to classical approaches in ecology or microbiome.

Keywords : Count data · Pólya distribution · Splitting model · Multivariate Analysis · Joint Species Distribution Model.

Introduction

Modeling multivariate count data is crucial in many applied fields. In ecology, jointly modeling species distribution according for environmental factors is of primary importance for predicting the impact of climate changes at the ecosystem scale [Ovaskainen and Soininen, 2011; Warton et al., 2015; Bry et al., 2020]. Similar challenges arise in the microbiome context, where understanding microbial community composition may help in defining individual healthcare strategies [Chen and Li, 2013; Wang and Zhao, 2017; Tang et al., 2018], or in econometric analysis to evaluate the number of transactions between various companies [Winkelmann, 2008]. Finding the appropriate model remains challenging. In particular, some data set may exhibit simultaneously positive or negative correlations between different pairs of variables. An ideal model should be flexible enough to take into account such a correlation structure, while remaining simple for inference and interpretation. Further consideration should be given to marginal distributions, which may be overdispersed due to an excess of zeros and/or extreme values present in the sample.

Given these constraints, several models have been proposed, such as the multivariate generalized Waring distribution [Xekalaki, 1986], the discrete Schur-constant model [Castañer et al., 2015], and the negative multinomial. An essential feature of these examples is their representation. Indeed, as presented by Jones and Marchand [2019] and Peyhardi

et al. [2021], these models belong to a large class of distributions where each can be expressed as a composition of a univariate discrete distribution and a singular multivariate distribution. Precisely, Jones and Marchand [2019] proposed the *sums and shares* model where $\mathbf{Y} = (Y_1, \ldots, Y_J) \in \mathbb{N}^J$ is such that the distribution of $\mathbf{Y}$ given $\sum_{j=1}^{J} Y_j = n$ is Dirichlet-multinomial, and the distribution of $\sum_{j=1}^{J} Y_j$ is negative binomial. Peyhardi et al. [2021] generalized those results using singular distributions with certain properties (e.g. multivariate Pólya distributions introduced by Eggenberger and Pólya [1923]), and an arbitrary discrete univariate distribution on the sum. Intuitively, their models can be interpreted as the random sharing of a univariate random variable into J categories. This simple stochastic representation where the sum of $\mathbf{Y} \in \mathbb{N}^J$ is randomly split by a Pólya distribution is called the *Pólya Splitting* distribution. This class emerges naturally as stationary distributions of a multivariate birth–death process under extended neutral theory [Peyhardi et al., 2024], possessing tractable univariate and multivariate marginals. Its dispersion is well understood, as is the dependence structure. However, the latter is quite restricted since all pairwise correlations must have identical signs.

Many applications consider only the singular multivariate distribution. This is particularly true in biology, where RNA-sequences are studied. In this field, the Dirichlet-multinomial and its many generalizations are widely utilized [e.g. Chen and Li, 2013]. The generalized Dirichlet-multinomial, introduced by Connor and Mosimann [1969], is considered by Tang and Chen [2018] with a focus on zero-inflation. Another example is the Dirichlet-tree multinomial model proposed by Dennis [1991] and employed by Wang and Zhao [2017] for gut microorganisms. These examples also have an interesting representation. Indeed, for $\mathbf{Y} \in \mathbb{N}^J$ such that $\sum_{j=1}^{J} Y_j = n$, each distribution can be interpreted as a stochastic process where the total is recursively split by multiple Dirichlet-multinomial distributions. Such a process can be represented by a tree structure where each node is distributed conditionally as Dirichlet-multinomial, and the leaves are the marginals Y_j.

Wang and Zhao [2017] justified their application of the Dirichlet-tree multinomial with a phylogenetic tree structure. This tree-like structure of the distribution enables various sign of correlation, but is less flexible since $\sum_{j=1}^{J} Y_j$ is fixed and not random. In fact, as we will show in this work, such a constraint has a significant impact on the same correlation structure, but also on its marginals.

Aitchison and Ho [1989] proposed the multivariate Poisson-lognormal as a solution to provide flexibility for modeling both correlations and marginals. This model exhibits a diverse correlation structure due to its underlying multivariate lognormal distribution as a latent variable. However, from an application perspective, it is important to note that these dependencies represent the latent space rather than the observations. Specifically, a null correlation does not imply independence among observations. A compelling model would be able to be as flexible as the Poisson-lognormal while accurately capturing the true dependencies of the data. This can be achieved by combining the Pólya Splitting approach and the tree structure of the Dirichlet-tree multinomial.

In this article, we propose a new class of multivariate discrete distributions named *Tree Pólya Splitting*, which combines a univariate random variable with a tree singular distribution where each node is associated with a Pólya split. As it will be demonstrated, this simple modification of the Pólya Splitting enables a diverse correlation structure with genuine dependencies and overdispersed marginals. This composition of sum and tree Splitting also allows for a straightforward inference approach where each component is estimated independently. Since this new model is a generalization of Peyhardi et al. [2021], we present various properties of the Tree Pólya Splitting and compare them to those of the Pólya Splitting. This paper is organized as follows. Section 1 introduces notations and basic results of Pólya Splitting that are used throughout. We also provide new results concerning the dispersion of marginal distributions and bounds for correlations. Section 2 is dedicated to the Tree Pólya Splitting distributions. We first define the tree

structure and then the associated distribution. Following this, properties of marginal distributions, factorial moments, and covariance/correlation are presented. A detailed study is carried out for each property with the help of a running example. Finally, in Section 3, we present a simple application of our new model to the Trichoptera data set provided by Usseglio-Polatera and Auda [1987] and compare it to the Poisson-lognormal. We also briefly explore how the observed data can inform us on the underlying tree structure. All proofs of properties and propositions presented in this paper are given in the Appendix.

1 Pólya Splitting distributions

This section presents notations, definitions, and properties of the Pólya Splitting distribution used throughout the paper. Precisely, the marginal distributions, factorial moments, and Pearson correlation structure of the Pólya Splitting model will be presented. These will be used as building blocks for our generalization of the model. This section also expands upon previous work by further analyzing the behavior of the covariance/correlation, and presenting the importance of dispersion for this discrete multivariate model. We refer to Jones and Marchand [2019], Peyhardi et al. [2021], and Peyhardi [2023] for further details.

1.1 Definitions and notations

Vectors and scalars will be denoted by bold and plain letters, respectively. For a vector $\mathbf{y} = (y_1, \dots, y_J)$, the sum of its components is denoted by $|\mathbf{y}| = \sum_{j=1}^{J} y_j$. For $\mathcal{J} \subset \{1, \dots, J\}$ a subset of indexes, we define $\mathbf{y}_{\mathcal{J}}$ and its complement $\mathbf{y}_{-\mathcal{J}}$ as $\mathbf{y}_{\mathcal{J}} = (y_j)_{j \in \mathcal{J}}$ and $\mathbf{y}_{-\mathcal{J}} = (y_j)_{j \in \mathcal{J}^c}$. Also, any binary operation between vectors is taken component-wise. The *discrete simplex* will be denoted by $\Delta_n := \{\mathbf{y} \in \mathbb{N}^J : |\mathbf{y}| = n\}$. For $\theta \in \mathbb{R}_+$, $c \in \{-1, 0, 1\}$ and

$n \in \mathbb{N}$, the function $(\theta)_{(n,c)}$ denotes the *generalized factorial* given by

$$(\theta)_{(n,c)} \begin{cases} 1 & \text{if } n = 0 \\ \theta(\theta + c) \dots (\theta + (n-1)c) & \text{if } n \geq 1. \end{cases} \tag{1}$$

If $c = 0$, then $(\theta)_{(n,0)} = \theta^n$, while $c = -1$ and $c = 1$ correspond respectively to the *falling* and *rising* factorial. The latter will be denoted by the Pochhammer symbol $(\theta)_n := (\theta)_{(n,1)}$. We also have $(\theta)_x = \Gamma(\theta + x)/\Gamma(\theta)$ for $\theta, x \in \mathbb{R}_+$. Furthermore, the falling factorial is related to the rising factorial as follows : $(\theta)_{(n,-1)} = (-1)^n(-\theta)_n$. Finally, for any $\boldsymbol{\theta} \in \mathbb{R}_+^J$, $\mathbf{r} \in \mathbb{N}^J$ and $n \in \mathbb{N}$, let us denote $(\boldsymbol{\theta})_{\mathbf{r}} := \prod_{j=1}^J (\theta_j)_{r_j}$ and $(\boldsymbol{\theta})_n := \prod_{j=1}^J (\theta_j)_n$.

A random variable $\mathbf{Y} \in \Delta_n$ is *Pólya* distributed if its probability mass function (p.m.f.) is given by

$$\mathbf{p}_{|\mathbf{Y}|=n}(\mathbf{y}) = \frac{n!}{(|\boldsymbol{\theta}|)_{(n,c)}} \prod_{j=1}^J \frac{(\theta_j)_{(y_j,c)}}{y_j!}, \tag{2}$$

for $c \in \{-1, 0, 1\}$ and parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_J)$ [Eggenberger and Pólya, 1923; Johnson et al., 1997]. Such a distribution will be denoted by $\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta})$. The following distributions are retrieved : the hypergeometric distribution $\mathcal{H}_{\Delta_n}(\boldsymbol{\theta})$ ($c = -1$), the multinomial $\mathcal{M}_{\Delta_n}(\boldsymbol{\theta})$ ($c = 0$), and the Dirichlet-multinomial $\mathcal{DM}_{\Delta_n}(\boldsymbol{\theta})$ ($c = 1$). In order to have an adequate distribution on Δ_n, the allowable values of $\boldsymbol{\theta}$ are the following : $\boldsymbol{\theta} \in \mathbb{R}_+^J$ for $c \in \{0, 1\}$, and $\boldsymbol{\theta} \in \mathbb{N}_+^J$ such that $|\boldsymbol{\theta}| \geq n$ for $c = -1$. Additionally, since the Pólya distribution is singular, i.e. its support has $J - 1$ degrees of freedom, the univariate version of the Pólya distribution will be denoted by $\mathcal{P}_n^{[c]}(\theta, \tau)$ when $J = 2$. Combining this distribution with the hypothesis that $|\mathbf{Y}| \sim \mathcal{L}(\boldsymbol{\psi})$, an univariate discrete distribution, we have the *Pólya Splitting distribution* defined as follows.

Definition 1. (Pólya Splitting distribution) A random vector $\mathbf{Y} = (Y_1, \dots, Y_J) \in \mathbb{N}^J$ follows a Pólya Splitting distribution with parameters c, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and generating distribution $\mathcal{L}(\boldsymbol{\psi})$ if $|\mathbf{Y}| \sim \mathcal{L}(\boldsymbol{\psi})$ and $\mathbf{Y} \,|\, |\mathbf{Y}| = n \sim \mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta})$. This decomposition is denoted by

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

Its p.m.f. is given by

$$\mathbf{p}(\mathbf{y}) = \mathbf{p}(|\mathbf{Y}| = n)\left[\frac{n!}{(|\boldsymbol{\theta}|)_{(n,c)}}\prod_{j=1}^{J}\frac{(\theta_j)_{(y_j,c)}}{y_j!}\right], \tag{3}$$

with $n = |\mathbf{y}|$ and $\mathbf{p}(|\mathbf{Y}| = n)$ the p.m.f. of $\mathcal{L}(\boldsymbol{\psi})$.

Before proceeding, the case $c = -1$ needs to be carefully analyzed. Indeed, the restriction on $|\boldsymbol{\theta}| \geq n$ is stated for n fixed. However, in a Pólya Splitting model, this value is random. Therefore, it is required that the support of $\mathcal{L}(\boldsymbol{\psi})$ be finite with upper bound value $m \in \mathbb{N}_+$, in which case the Pólya Splitting distribution for $c = -1$ is well defined if $|\boldsymbol{\theta}| \geq m$.

1.2 Marginal distributions and factorial moments

The marginals of the Pólya Splitting distribution are themselves Pólya Splitting. Indeed, Peyhardi et al. [2021] show that for $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, the marginal distribution of Y_j is given by

$$Y_j \sim \mathcal{P}^{[c]}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi}). \tag{4}$$

Notice that the univariate distribution $\mathcal{L}(\boldsymbol{\psi})$ in (4) is, in a sense, "damaged" by the univariate Pólya distribution. Such a composition is, in fact, similar to the binomial thinning operator studied in Rao [1965] and used in the time series model proposed by Joe [1996]. For more details concerning this type of operator, see Davis et al. [2021]. Peyhardi [2023] presents three general families of distributions that are stable, i.e. distributions $\mathcal{L}$ such that

$$\mathcal{L}(\tilde{\boldsymbol{\psi}}) = \mathcal{P}^{[c]}_n(\theta, \tau) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$$

where $\tilde{\boldsymbol{\psi}}$ are updated parameters of $\boldsymbol{\psi}$. One of such family consists of *power series distribution*, denoted by $\mathcal{PS}^{[c]}(\theta, \alpha)$, with p.m.f. given by

$$\mathbf{p}(y) \propto \frac{\alpha^y}{y!}(\theta)_{(y,c)}, \tag{5}$$

where the values of the parameters α, θ, and the support of Y depend on c. For each value of $c \in \{-1, 0, 1\}$, the corresponding distributions are given by the binomial, Poisson, and negative binomial distributions respectively. See Table 1 for each distribution represented in terms of α, θ and c.

Distributions	Parameters	Support	P.m.f.
$\mathcal{B}_\theta(\alpha)$	$\theta \in \mathbb{N}_+,\ \alpha \in \mathbb{R}_+$	$y \in \Delta_\theta$	$\binom{\theta}{y}\left(\frac{\alpha}{\alpha+1}\right)^y\left(\frac{1}{\alpha+1}\right)^{\theta-y}$
$\mathcal{P}(\alpha\theta)$	$(\theta, \alpha) \in \mathbb{R}_+^2$	$y \in \mathbb{N}$	$\frac{(\alpha\theta)^n}{n!}e^{-\alpha\theta}$
$\mathcal{NB}(\theta, \alpha)$	$\theta \in \mathbb{R}_+,\ \alpha \in (0,1)$	$y \in \mathbb{N}$	$\frac{(\theta)_y}{y!}\alpha^y(1-\alpha)^\theta$

TABLE 1 – Power distributions for $c = -1$, 0 and 1 respectively

For Pólya Splitting, the power series $\mathcal{PS}^{[c]}(\theta, \alpha)$ are the only distributions which allow the marginals to be independent. Indeed, Peyhardi [2023] shows that for $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, the Y_j's are independent if and only if $\mathcal{L}(\boldsymbol{\psi}) = \mathcal{PS}^{[c]}(|\boldsymbol{\theta}|, \alpha)$. Finally, the factorial moments need to be defined. For a univariate random variable, the r-th factorial moment of $Y \in \mathbb{N}$ is the expected value of the r-th falling factorial, i.e.

$$\mathrm{E}[Y(Y-1)\cdots(Y-r+1)] = (-1)^r\mathrm{E}[(-Y)_r].$$

Similarly for $\mathbf{r} = (r_1, \ldots, r_J) \in \mathbb{N}^J$, the multivariate factorial moment of $\mathbf{Y} \in \mathbb{N}^J$ is the expectation $(-1)^{|\mathbf{r}|}\mathrm{E}\left[(\mathbf{Y})_{\mathbf{r}}\right]$. For a fixed n, the factorial moments of the Pólya distribution can be obtained using the Chu-Vandermonde identity [Johnson et al., 1997], which leads to the following.

Property 1. For $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and $\mathbf{r} \in \mathbb{N}^J$, the multivariate factorial moments are given by

$$(-1)^{|\mathbf{r}|}\mathrm{E}\left[(-\mathbf{Y})_{\mathbf{r}}\right] = \frac{\mu_{|\mathbf{r}|}}{(|\boldsymbol{\theta}|)_{(|\mathbf{r}|,c)}}\prod_{j=1}^{J}(\theta_j)_{(r_j,c)},$$

where μ_r is the r-th factorial moment of $\mathcal{L}(\boldsymbol{\psi})$.

1.3 Covariance and dispersion

Using Property 1, the covariance between Y_i and Y_j for $i \neq j$ in the Pólya Splitting distribution is given by

$$\mathrm{Cov}(Y_i, Y_j) = \frac{\theta_i \theta_j}{|\boldsymbol{\theta}|^2 (|\boldsymbol{\theta}| + c)} \left[(\mu_2 - \mu_1^2) |\boldsymbol{\theta}| - c\mu_1^2 \right]. \tag{6}$$

Here, we are particularly interested in the signs of the covariances. Clearly, the sign of (6) is related to the hyperplane defined by $(\mu_2 - \mu_1^2)|\boldsymbol{\theta}| = c\mu_1^2$ and separates the parameter values $\boldsymbol{\theta}$ into regions of negative, positive, and null covariance. Observe as well that the covariances have the same sign for all pair (Y_i, Y_j). Additionally, note that $\mu_2 - \mu_1^2 = \mathrm{Var}[|\mathbf{Y}|] - \mathrm{E}[|\mathbf{Y}|]$. This determines what type of dispersion the distribution $\mathcal{L}(\boldsymbol{\psi})$ has. There are three possible situations, $\mathcal{L}$ is *underdispersed* if $\mu_2 - \mu_1^2 < 0$, *overdispersed* if $\mu_2 - \mu_1^2 > 0$, or has a *null dispersion* if $\mu_2 - \mu_1^2 = 0$. Each type of dispersion and value c of the Pólya yields different situations. For $c = 0$, the sign of covariance is simply determined by the dispersion of $\mathcal{L}$. In the case of Dirichlet-multinomial Splitting (i.e. $c = 1$), then if $\mathcal{L}$ is underdispersed or has null dispersion, the sign is always negative. However, if $\mathcal{L}$ is overdispersed, the sign of covariance is negative, null or positive if and only if $|\boldsymbol{\theta}|$ is less, equal or greater than $\mu_1^2/(\mu_2 - \mu_1^2)$ respectively. A similar analysis for $c = -1$ can be made using the restriction on $\boldsymbol{\theta}$.

Since the dispersion of the distribution $\mathcal{L}$ is relevant for the covariance sign, it will be useful for our model to understand how this dispersion is preserved in the marginals. For example, if $\mathcal{L}$ is overdispersed, does it imply that the marginals are necessary overdispersed ? We have the following.

Property 2. For $\mathbf{Y} \sim \mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, then :

— If $c = 0$, the marginals have the same type of dispersion as $\mathcal{L}$;

— If $c = 1$ and $\mathcal{L}$ has null or positive dispersion, then the marginals are overdispersed ;

— If $c = -1$ and $\mathcal{L}$ has null or negative dispersion, then the marginals are underdispersed.

Property 2 implies that the type of dispersion is preserved for $c = 0$, but can change for other values. For example, if $\mathcal{L}$ is underdispersed and $c = 1$, then it is possible to have different dispersion at the marginals. In this case, the dispersion is determined by the values of $\boldsymbol{\theta}$.

1.4 Pearson correlation structure

An interesting way to formulate the correlation between Y_i and Y_j is to use the relation between factorial moments of $\mathcal{L}$ and the marginals. Indeed, by Property 1 and any pair $i \neq j$, then

$$\mathrm{E}[(-Y_j)_r] = \frac{(\theta_j)_{(r,c)}}{(\theta_i)_{(r,c)}} \mathrm{E}[(-Y_i)_r].$$

Using this identity, the quantities $\mathrm{Cov}(Y_i, Y_j)$ and $\mathrm{Var}[Y_j]$ can be expressed in terms of Y_i so that the following property can be proved.

Property 3. For $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and $i \neq j$, then the Pearson correlation coefficient is given by

$$\mathrm{Corr}(Y_i, Y_j) = \sqrt{\frac{\theta_i \theta_j}{(\theta_i + c)(\theta_j + c)}} \frac{(1 - M_i)}{\sqrt{1 - \left(\frac{\theta_j - \theta_i}{\theta_j + c}\right) M_i}},$$

where

$$M_i = \frac{\mathrm{E}[Y_i]}{\mathrm{Var}[Y_i]} \left(1 + \frac{c}{\theta_i} \mathrm{E}[Y_i]\right) = \frac{\mu_1 \left(1 + \frac{c}{|\boldsymbol{\theta}|} \mu_1\right)}{\mu_2 \left(\frac{\theta_i + c}{|\boldsymbol{\theta}| + c}\right) + \mu_1 \left(1 - \mu_1 \frac{\theta_i}{|\boldsymbol{\theta}|}\right)},$$

and μ_r the r-th factorial moment of $\mathcal{L}$.

Jones and Marchand [2019] showed that for any distribution $\mathcal{L}$, $\mathrm{Corr}(Y_i, Y_j) < 1/2$ when $\boldsymbol{\theta} = \mathbf{1}$, the unit vector, and $c = 1$. With Property 3, we can generalize their result to other values of $\boldsymbol{\theta}$.

Property 4. For $\mathbf{Y} \sim \mathcal{DM}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and $i \neq j$, then the correlation is such that

$$\text{Corr}(Y_i, Y_j) < \sqrt{\frac{\theta_i \theta_j}{(\theta_i + 1)(\theta_j + 1)}} \ .$$

Notice that this bound is not sharp. Interestingly, this bound is equal to the geometric mean of $\theta_i/(\theta_i + 1)$ and $\theta_j/(\theta_j + 1)$.

2 Tree Pólya Splitting Distribution

In this section, we first present the notations and definitions of rooted trees inspired by Tang et al. [2018]. Following this, we define the Tree Pólya Splitting distribution and present similar properties of the previous section. As we shall see, all previous properties are simply particular cases of our generalization. A running example is used throughout this section to illustrate and explore further these results. In particular, we are able to obtain a new marginal p.m.f. that generalizes Jones and Marchand [2019] results. We also show how the correlations of the Tree Pólya can indeed take various signs.

2.1 Definitions and Notations

Let $\mathfrak{T} = (\mathfrak{N}, \mathfrak{E})$ be a undirected graph with nodes $\mathcal{N}$ and edges $\mathcal{E}$. $\mathfrak{T}$ is an *undirected tree* if it is connected, i.e. there is a path of edges between every pair of nodes in the graph, and acyclic, i.e. the graph contains no cycle. Furthermore, $\mathfrak{T}$ is a *rooted tree*, or *directed tree*, if it is an undirected tree with a fixed node named root. Fixing such a node gives $\mathfrak{T}$ an orientation from the root to the nodes called leaves. A *leaf* is a node such that only one edge is connected to it. In this instance, $\Omega \in \mathcal{N}$ will denote the root and $\mathfrak{L} \subseteq \mathcal{N}$ the subset of leaves. An *internal node* is any node that is not a leaf. The set of these nodes will be denoted by $\mathfrak{I}$. Finally, because a rooted tree has a direction, we can

establish a parent/child relation between nodes. For any internal node $A \in \mathfrak{I}$, its set of *children nodes*, denoted by $\mathfrak{C}_A$, contains any node directly connected to A in the opposite direction of the root. Such a set has elements $\mathfrak{C}_A = \{C_1, \ldots, C_{J_A}\}$ with $J_A \geq 2$ the number of children. Notice here that we assume that all internal nodes have at least two children. Similarly, for any node $A \in \mathfrak{I} \cup \mathfrak{L}$, its parent is the node directly connected to A in the direction of the root. It is denoted by $\mathscr{P}(A)$ and is such that : (i) $\mathscr{P}(\Omega) = \varnothing$, and (ii) $\mathscr{P}(C_i) = \mathscr{P}(C_j) = A$ for $C_i, C_j \in \mathfrak{C}_A$ with $i \neq j$, i.e. C_i and C_j are *sibling nodes*. Based on these definitions and notations, we are now able to define a specific type of rooted tree useful for our model.

Definition 2 (Partition tree)**.** A rooted tree $\mathfrak{T}$ is a partition tree if its root $\Omega = \{1, \ldots, J\}$, the leaves $\mathfrak{L} = \{\{1\}, \ldots, \{J\}\}$, and each sibling form a partition of their parent.

Finally, the notion of path between two nodes will be useful to understand various properties related to the Tree Pólya. Such a path is constructed through an iteration of the parent nodes from any leaf or internal node to another node.

Definition 3 (Path)**.** For a partition tree $\mathfrak{T}$, any $A \in \mathfrak{I} \cup \mathfrak{L}$, $B \in \mathfrak{I}$ such that $A \subset B$ and

$$\mathscr{P}_A^n := \underbrace{\mathscr{P}(\mathscr{P}(\ldots \mathscr{P}(A)) \ldots)}_{\text{n times}},$$

the n-th parent of node A with $\mathscr{P}_A^0 = A$, the path from A to B is defined by the ordered set

$$\text{Path}_A^B := \left(\mathscr{P}_A^0, \mathscr{P}_A^1, \mathscr{P}_A^2, \ldots, \mathscr{P}_A^K\right),$$

where K is such that $\mathscr{P}_A^K = B$. By convention, $A_n \in \text{Path}_A^B$ means that A_n is the n-th element of Path_A^B. Therefore, the element A_{n-1} for $n \geq 1$ should be interpreted as the child of A_n. Moreover, if $B = \Omega$, then $\text{Path}_A := \text{Path}_A^\Omega$.

With these definitions, the structure of the partition tree can be fully described and used. For our running example, we will use the partition tree presented in Figure 1. In

this example, $\Omega = \{1, \ldots, 10\}$, $A = \{4, \ldots, 10\}$ is an internal node with children nodes $\mathfrak{C}_A = \{\{4,5\}, \{6,7\}, \{8,9,10\}\}$, $\mathscr{P}(A) = \Omega$, and the path between the leaf $\{9\}$ and A is given by $\text{Path}_{\{4\}}^{A} = (\{9\}, \{9,10\}, \{8,9,10\}, A)$.

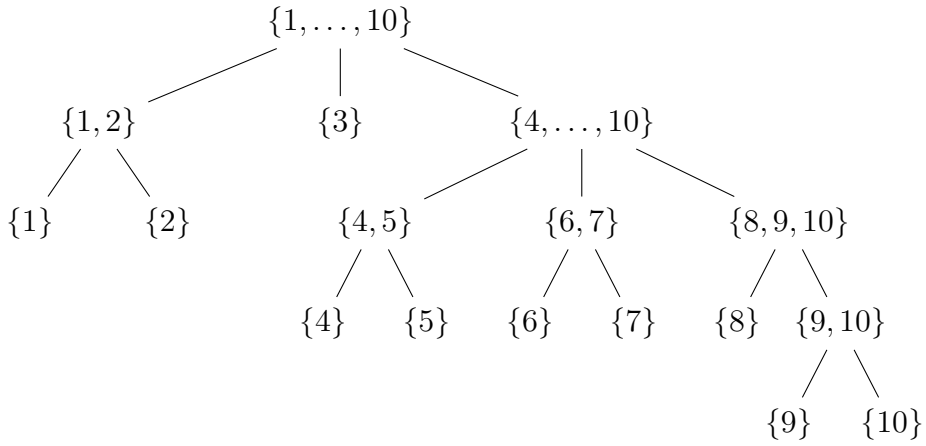

FIGURE 1 – Example of a partition tree with 10 leaves.

It is now possible to generalize the Pólya Splitting distribution with the partition tree as follows. For $\mathbf{Y} \in \mathbb{N}^J$, $|\mathbf{Y}|$ can be split by a Pólya into subsums which are again split until each marginal Y_j is obtained. It is assumed that these divisions are fixed by the model, i.e. we know which Y_j are used in each subsum. This approach allows to create various clusters of $\mathbf{Y}$ with different dependence structure or marginals. Since the order of divisions is fixed, this new distribution can be represented and studied with the partition tree $\mathfrak{T}$. Indeed, the internal nodes $\mathfrak{I}$ determine all the subsums involved and the leaves $\mathfrak{L}$ represent all marginals. Moreover, since the divisions are independent Pólya distributions, the p.m.f. can be directly obtained. Thus we have the following definition.

Definition 4 (Tree Pólya Splitting distribution)**.** For a partition tree $\mathfrak{T}$, $\mathbf{Y} \in \mathbb{N}^J$ is said to follow a *Tree Pólya Splitting* distribution if, for each node $A \in \mathfrak{I}$, the distribution of the subsums $\left(|\mathbf{Y}_{C_1}|, \ldots, |\mathbf{Y}_{C_{J_A}}|\right)$ given $\mathbf{Y}_{\mathbf{A}} = n$ is $\mathcal{P}_n^{[c_A]}(\boldsymbol{\theta}_A)$, with $\boldsymbol{\theta}_A = \{\theta_C\}_{C \in \mathfrak{C}_A}$ and c_A depending on A. Such a distribution is denoted by $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ with

$\boldsymbol{\theta} = \{\boldsymbol{\theta}_A\}_{A\in\mathfrak{J}}$, $\mathbf{c} = \{c_A\}_{A\in\mathfrak{J}}$ and p.m.f.

$$\mathbf{p}(\mathbf{y}) = \mathbf{p}(|\mathbf{Y}| = n) \prod_{A\in\mathfrak{J}} \frac{n_A!}{(|\boldsymbol{\theta}_A|)_{(n_A,c_A)}} \prod_{C\in\mathfrak{C}_A} \frac{(\theta_C)_{(n_C,c_A)}}{n_C!}, \tag{7}$$

where $n_A := |\mathbf{y}_A|$ for any node A and $n_\Omega := n$ by definition.

Notice that if $\mathfrak{J} = \Omega$ in Definition 4, then the basic Pólya Splitting p.m.f. (3) is recovered in (7). Moreover, the largest number of parameters needed is attained for the binary tree, i.e., each internal node has two children. Therefore, depending on the type of Splittings, $\mathcal{TP}_{\Delta_n}(\mathfrak{T};\boldsymbol{\theta},\mathbf{c})$ number of parameters varies between $|\boldsymbol{\psi}| + (J-1)$ and $|\boldsymbol{\psi}| + 2(J-1)$. Now, using the partition tree in Figure 1, Figure 2 presents the model representation of our running example where all internal nodes are either a multinomial ($\mathcal{M}$) or a Dirichlet-multinomial ($\mathcal{DM}$), and each edge is associated to a parameter of the given Pólya. Moreover, the distribution of $|\mathbf{Y}|$ is indicated at the top of the tree.

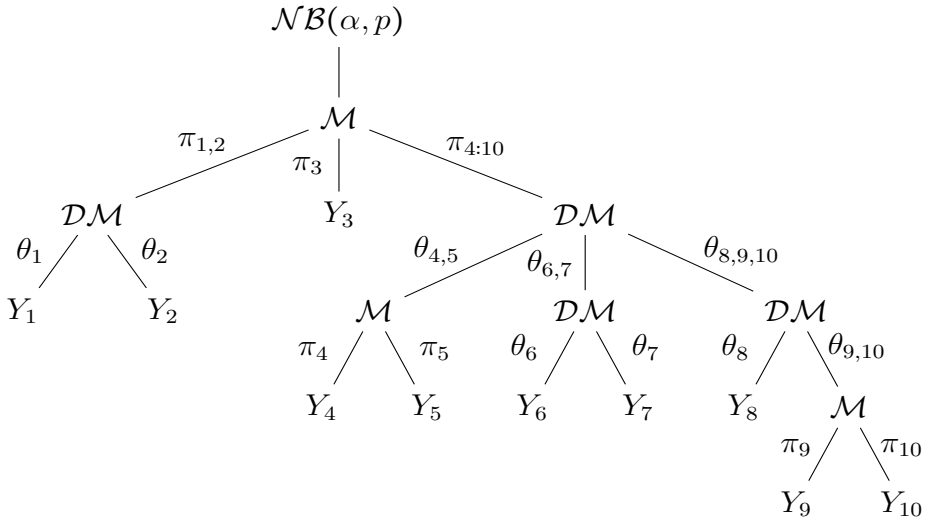

FIGURE 2 – Example of Tree Pólya Splitting distribution based on the partition tree in Figure 1 and $\mathcal{L} = \mathcal{NB}(\alpha, p)$

.

Just like Pólya Splitting, several known distributions are particular cases of Tree Pólya Splitting. As previously indicated, the Pólya Splitting itself is a trivial case. For a fixed

value of $|\mathbf{Y}|$, i.e. the total follows a Dirac, the generalized Dirichlet-multinomial can be directly retrieved [Connor and Mosimann, 1969]. Indeed, the tree $\mathfrak{T}$ is such that the elements of $\mathfrak{I}$ are given by $A_j = \{j, \ldots, J\}$ for all $j \in \{1, \ldots, J\}$ and their set of children is given by $\mathfrak{C}_{A_j} = \{\{j\}, \{j+1, \ldots, J\}\}$. For each node, a Dirichlet-multinomial is used and can be represented by a binary cascade tree (Figure 3). Similarly, the Dirichlet-tree multinomial [Dennis, 1991] use a more general tree structure where each internal node are again distributed as Dirichlet-multinomial. In all of the above examples, it is important to keep in mind that the value $|\mathbf{Y}|$ is fixed and not random. As we will demonstrate, the added randomness of $|\mathbf{Y}|$ can have a significant impact on the covariance structure. Additionally, the order of the marginals is important for any Tree Pólya Splitting. Indeed, for $\mathbf{Y} = (Y_1, \ldots, Y_J)$ and $\tilde{\mathbf{Y}} = (\mathbf{Y}_{\sigma(1)}, \ldots, \mathbf{Y}_{\sigma(J)})$ with $\sigma(\cdot)$ a non-identity permutation, the p.m.f. of the tree Splitting is such that $\mathbf{p}(\mathbf{y}) \neq \mathbf{p}(\tilde{\mathbf{y}})$. Therefore, the order of the leaves $\mathfrak{L}$ in the tree should always be kept in mind.

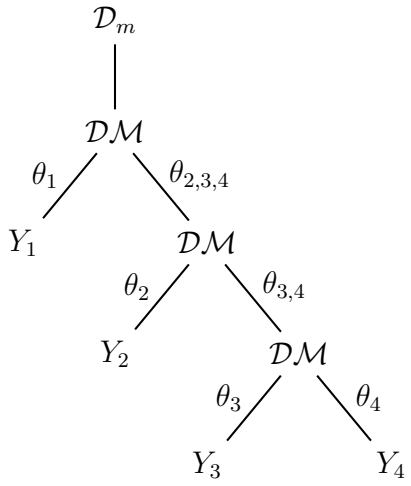

FIGURE 3 – Generalized Dirichlet-multinomial distribution represented by a Tree Pólya Splitting model with $\mathcal{L}(\boldsymbol{\psi}) = \mathcal{D}_m$, the Dirac distribution at m.

2.2 Properties

We expand here on properties of Tree Pólya Splitting distributions, which extend those of the Pólya Splitting. To illustrate them, we use our running example presented in Figure 2, where the total follows a negative binomial distribution, and each internal node can either be a multinomial or a Dirichlet-multinomial. To understand the impact of $\mathcal{L}$, we compare our example to the same Tree Pólya Splitting but with a fixed total.

2.2.1 Marginals

Since the Tree Pólya Splitting is simply an iteration of different Splittings throughout the partition tree, the marginal should have a similar form as equation (4). Intuitively, because any marginal Y_j is represented by a leaf in the partition tree, its path to the root must dictate the form of its distribution. The following proposition shows it is indeed the case.

Proposition 1 (Univariate marginal for a leaf)**.** For $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, the distribution of Y_j is given by

$$Y_j \sim \bigwedge_{k=1}^{K} \mathcal{P}_{n_k}^{[c_{A_k}]}\left(\theta_{A_{k-1}}, |\boldsymbol{\theta}_{A_k \setminus A_{k-1}}|\right) \underset{n_K}{\wedge} \mathcal{L}(\boldsymbol{\psi}), \tag{8}$$

where $A_k \in \mathrm{Path}_{\{j\}}$, $\boldsymbol{\theta}_{A_k \setminus A_{k-1}}$ is the set of parameters at node A_k minus the parameter $\theta_{A_{k-1}}$, and

$$\bigwedge_{k=1}^{K} \mathcal{P}_{n_k}^{[c_{A_k}]}\left(\theta_{A_{k-1}}, |\boldsymbol{\theta}_{A_k \setminus A_{k-1}}|\right) := \mathcal{P}_{n_1}^{[c_{A_1}]}\left(\theta_{A_0}, |\boldsymbol{\theta}_{A_1 \setminus A_0}|\right) \underset{n_1}{\wedge} \quad \ldots \underset{n_{K-1}}{\wedge} \mathcal{P}_{n_K}^{[c_{A_K}]}\left(\theta_{A_{K-1}}, |\boldsymbol{\theta}_{A_K \setminus A_{K-1}}|\right).$$

Any partial sum must have a similar distribution since it is represented by an internal node. Therefore, there is a path from the root to the latter. The next result follows directly from Proposition 1.

Proposition 2 (Univariate marginal for a partial sum in $\mathfrak{I}$)**.** For $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and an internal node $A \in \mathfrak{I}$, the marginal distribution of $|\mathbf{Y}_A|$ is given by (8) but with $A_k \in \text{Path}_A$. If $A = \Omega$, then $K = 0$ and $|\mathbf{Y}| \sim \mathcal{L}(\boldsymbol{\psi})$.

Finally, Proposition 2 can be used to obtain multivariate marginal distributions that are consistent with the whole tree structure.

Proposition 3 (Multivariate marginal of a subtree)**.** For $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and an internal node $A \in \mathfrak{I}$, the multivariate marginal distribution $\mathbf{Y}_A$ is again Tree Pólya Splitting, i.e.

$$\mathbf{Y}_A \sim \mathcal{TP}_{\Delta_n}(\tilde{\mathfrak{T}}; \tilde{\boldsymbol{\theta}}, \tilde{\mathbf{c}}) \underset{n}{\wedge} |\mathbf{Y}_A|,$$

where the distribution $|\mathbf{Y}_A|$ is given by Proposition 2, $\tilde{\mathfrak{T}}$ is the subtree with root A, $\tilde{\mathfrak{I}} = (B \in \mathfrak{I} : B \subseteq A)$ and $\tilde{\mathfrak{L}} = (\{j\} \in \mathfrak{L} : \{j\} \subseteq A)$. Finally, $\tilde{\boldsymbol{\theta}} = \{\boldsymbol{\theta}_A\}_{A \in \tilde{\mathfrak{I}}}$ and $\tilde{\mathbf{c}} = \{\mathbf{c}_A\}_{A \in \tilde{\mathfrak{I}}}$ are the parameters involved in the subtree.

Running example

As an example, consider the distribution of Y_6 in Figure 2. Using Proposition 1, the marginal is given by

$$Y_6 \sim \mathcal{BB}_{n_1}(\theta_6, \theta_7) \underset{n_1}{\wedge} \mathcal{BB}_{n_2}(\theta_{6,7}, \theta_{4,5} + \theta_{8,9,10}) \underset{n_2}{\wedge} \mathcal{B}_{n_3}(\pi_{4:10}) \underset{n_3}{\wedge} \mathcal{NB}(\alpha, p) \tag{9}$$

as represented by the path in Figure 4.

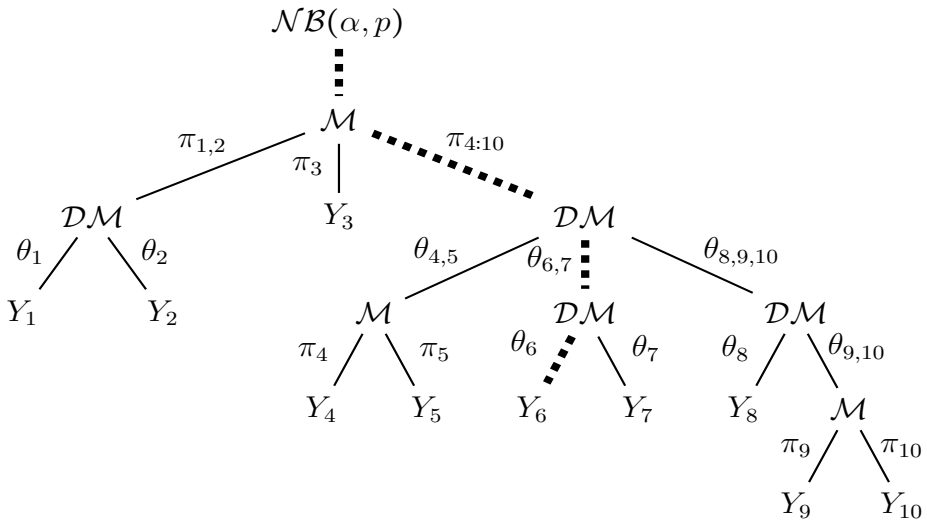

FIGURE 4 – Path representation of the marginal Y_6 in our running example.

In fact, equation (9) can be expressed with a composition of only beta-binomial distributions. Indeed, if the marginal distribution is the composition of binomial, beta-binomial and negative binomial distributions, then all the binomial distributions can be "absorbed" in the negative binomial by the following result.

Proposition 4. Suppose for K Pólya distributions there are M cases with $c_k = 1$ and parameters $\alpha_k, \beta_k \in \mathbb{R}_+$, and $K - M$ cases with $c_k = 0$ and parameters $\pi_k \in (0, 1)$. Then

$$\bigwedge_{k=1}^{K} \mathcal{P}_{n_k}^{[c_k]}(\theta_k, \tau_k) \underset{n_K}{\wedge} \mathcal{NB}(\alpha, p) = \left[\bigwedge_{m=1}^{M} \mathcal{BB}_{n_m}(\alpha_m, \beta_m)\right] \underset{n_M}{\wedge} \mathcal{NB}\left(\alpha, \frac{p\gamma}{1 - p(1-\gamma)}\right),$$

where $\gamma = \prod_{k=1}^{K-M} \pi_k$ and (θ_k, τ_k) is given by $(\pi_k, 1-\pi_k)$ or (α_k, β_k) whether $c_k = 0$ or $c_k = 1$ respectively for all k.

Therefore, to obtain the p.m.f. of (9), or any marginal in Figure 4, it is sufficient to study the p.m.f. of the general composition

$$X \sim \left[\bigwedge_{k=1}^{K} \mathcal{BB}_{n_k}(\alpha_k, \beta_k)\right] \underset{n_K}{\wedge} \mathcal{NB}(\alpha, p). \tag{10}$$

For $K = 1$, Jones and Marchand [2019] showed that the p.m.f. of (10) is given by

$$\mathbf{p}(n) = (1-p)^{\alpha} \frac{(\alpha)_n (\alpha_1)_n}{(\alpha_1+\beta_1)_n} \frac{p^n}{n!} {}_2F_1\left[\begin{matrix} \alpha+n,\ \beta_1 \\ \alpha_1+\beta_1+n \end{matrix}; p\right];\ n \in \mathbb{N},$$

where ${}_pF_q\left[\begin{matrix}\mathbf{a}\\ \mathbf{b}\end{matrix}; z\right] = \sum_{k=0}^{\infty} \frac{(\mathbf{a})_k}{(\mathbf{b})_k} \frac{z^k}{k!}$ denotes the *generalized hypergeometric series* with $\mathbf{a} \in \mathbb{R}_+^p$, $\mathbf{b} \in \mathbb{R}_+^q$, and $p \in (0,1)$. This result can be generalized for any positive integer K. Using the distribution of the product of independent beta random variables [e.g. Tang and Gupta, 1984], we have the following proposition.

Proposition 5. Let $p \in (0,1)$, $\alpha > 0$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, $K \geq 2$ and $k \in \{2, \ldots, K\}$, define

$$\rho_i^{(k)} = \frac{\Gamma\left(\sum_{s=0}^{k-1} \beta_s + i\right)}{\Gamma\left(\sum_{s=0}^{k} \beta_s + i\right)} \sum_{s=0}^{i} \frac{\left(\alpha_k + \beta_k - \alpha_{k-1}\right)_s}{s!} \rho_{i-s}^{(k-1)}$$

with initial values $\rho_0^{(1)} = 1/\Gamma(\beta_1)$ and $\rho_i^{(1)} = 0$ otherwise. If X is distributed as in (10), then its p.m.f. is given by

$$\mathbf{p}(n) = (\boldsymbol{\alpha})_{\boldsymbol{\beta}} (1-p)^{\alpha} \frac{(\alpha)_n p^n}{n!} \sum_{i=0}^{\infty} \rho_i^{(K)} \mathrm{B}\big(|\boldsymbol{\beta}| + i, \alpha_K + n\big) {}_2F_1\left[\begin{matrix} \alpha+n,\ |\boldsymbol{\beta}|+i \\ \alpha_K + |\boldsymbol{\beta}| + n + i \end{matrix}; p\right],$$

where $\mathrm{B}(\cdot,\cdot)$ is the beta function. In particular, if $p \in (0, 1/2)$, then the p.m.f. is also given by

$$\mathbf{p}(n) = \frac{(\boldsymbol{\alpha})_n}{(\boldsymbol{\alpha}+\boldsymbol{\beta})_n} \frac{(\alpha)_n}{n!} \left(\frac{p}{1-p}\right)^n {}_{K+1}F_K\left[\begin{matrix} \alpha+n,\ \boldsymbol{\alpha}+n\mathbf{1} \\ \boldsymbol{\alpha}+\boldsymbol{\beta}+n\mathbf{1} \end{matrix}; \frac{p}{p-1}\right]$$

with $\mathbf{1}$ the unit vector.

From Property 2, we infer that all the marginals in our example are overdispersed. Indeed, since the negative binomial is overdispersed and the tree is composed of multinomial and Dirichlet-multinomial Splittings, all the subsums and leaves are overdispersed. For the marginal Y_6 in our example, the p.m.f. of (9) can be obtained using Propositions 4 and

5. As an example, we fixed the parameters to $\theta_6 = 3$, $\theta_7 = \theta_{6,7} = 1$, $\theta_{4,5} + \theta_{8,9,10} = 2$, and $\pi_{4:10} = 0.75$. For the total, we let $\alpha \in \{5, 20\}$ and $p = \{0.25, 0.45\}$ and present how the p.m.f. varies in Figure 5 by a bar chart. Each case is also compared to the initial $\mathcal{NB}(\alpha, p)$. We find that for all values of α and p, the probability of $Y_6 = 0$ drastically increases compared to the initial negative binomial. The whole negative binomial tail actually decreases to small values. Since the total gets iteratively damaged by the partition tree, this phenomenon is intuitively sound.

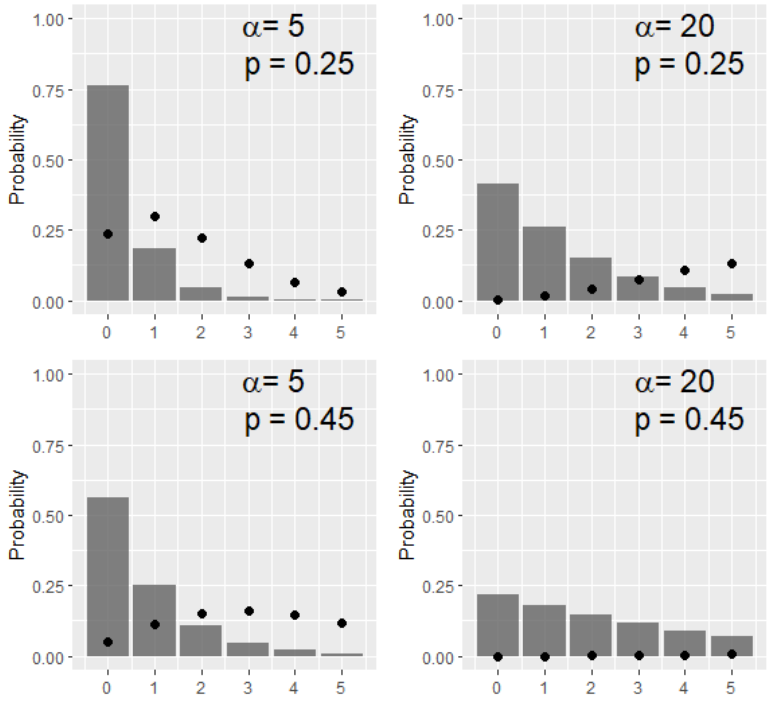

FIGURE 5 – P.m.f. of Y_6 (Bar) and $\mathcal{NB}(\alpha, p)$ (Points) with $\theta_6 = 3$, $\theta_7 = \theta_{6,7} = 1$, $\theta_{4,5}+\theta_{8,9,10} = 2$, $\pi_{4:10} = 0.75$, $\alpha \in \{5, 20\}$ and $p = \{0.25, 0.45\}$.

Now, if instead the total in the example is a fixed value m, i.e. a Dirac at m, then the Tree Pólya Splitting drastically changes. Indeed, the marginals are necessarily bounded and some marginals may be underdispersed or overdispersed by Property 2. Previously, the binomial splits were absorbed by the negative binomial in accordance to Proposition

4. However, because we replaced the negative binomial by a Dirac at m, we must find the p.m.f. of the random variable

$$X \sim \bigwedge_{k=1}^{K} \mathcal{P}_{n_k}^{[c_k]}(\theta_k, \tau_k) \underset{n_K}{\wedge} \mathcal{D}_m \tag{11}$$

where the parameters are given in Proposition 4. Using similar techniques, we prove the following.

Proposition 6. Let X be distributed as (11) with $m \in \mathbb{N}$ and suppose there are M Pólya with $c_k = 1$ and parameters $\alpha_k, \beta_k \in \mathbb{R}_+$, and $K - M$ Pólya with $c_k = 0$ and parameters $\pi_k \in (0,1)$. Then its p.m.f. is given by

$$\mathbf{p}(n) = \binom{m}{n} \frac{(\boldsymbol{\alpha})_n}{(\boldsymbol{\alpha}+\boldsymbol{\beta})_n} \gamma^n {}_{M+1}F_M\left[\begin{matrix} n-m,\ \boldsymbol{\alpha}+n\mathbf{1} \\ \boldsymbol{\alpha}+\boldsymbol{\beta}+n\mathbf{1} \end{matrix}; \gamma\right];\ n \in \{0,\dots,m\},$$

where $\gamma = \prod_{k=1}^{K-M} \pi_k$.

2.2.2 Factorial moments

Factorial moments of the Pólya Splitting distribution were determined by Property 1 to be the product of the J splits. In the same fashion, the factorial moments of the Tree Pólya should admit a similar product, but through its partition tree. The following proposition shows that it is indeed the case.

Proposition 7. For $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, $\mathbf{r} = (r_1, \dots, r_J) \in \mathbb{N}_+^J$, and by denoting $\mathbf{r}_A = (r_i)_{i \in A}$ for any $A \in \mathfrak{I} \cup \mathfrak{L}$, the factorial moments are given by

$$(-1)^{|\mathbf{r}|}\mathrm{E}[(-\mathbf{Y})_{\mathbf{r}}] = \mu_{|\mathbf{r}|} \prod_{A \in \mathcal{I}} \frac{\prod_{C \in \mathfrak{C}_A} (\theta_C)_{(|\mathbf{r}_C|, c_A)}}{(|\boldsymbol{\theta}_A|)_{(|\mathbf{r}_A|, c_A)}},$$

where $\mu_{|\mathbf{r}|}$ is the $|\mathbf{r}|$-th factorial moment of $\mathcal{L}$.

A direct corollary of Proposition 7 is the univariate factorial moments of any subsum in the tree. Here, instead of a product on all edges, the factorial moments are determined by the path from the root to the appropriate node or leaf.

Corollary 1. For $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, $r \in \mathbb{N}_+$, and any $A \in \mathfrak{I} \cup \mathfrak{L}$, the factorial moment of $|\mathbf{Y}_A|$ with $A_k \in \mathrm{Path}_A$ is given by

$$(-1)^r \mathrm{E}\left[-(|\mathbf{Y}_A|)_r\right] = \mu_r \prod_{k=1}^{K} \frac{(\theta_{A_{k-1}})_{(r,c_{A_k})}}{(|\boldsymbol{\theta}_{A_k}|)_{(r,c_{A_k})}},$$

where μ_r is the r-th factorial moment of $\mathcal{L}$.

Running example

The factorial moment of Y_6 in our example is determined by the same path as in Figure 4. Corollary 1 states that for $r \in \mathbb{N}_+$,

$$(-1)^r \mathrm{E}\left[(-Y_6)_r\right] = \mu_r \left(\pi_{4:10}^r \frac{(\theta_{6,7})_r}{(\theta_{4,5} + \theta_{6,7} + \theta_{8,9,10})_r} \frac{(\theta_6)_r}{(\theta_{4,5} + \theta_7)_r} \right).$$

In this case, the total distribution is given by $\mathcal{NB}(\alpha, p)$ with $\mu_r = (\alpha)_r \left(\frac{p}{1-p}\right)^r$. If instead the negative binomial is replaced by a Dirac at m, then $\mu_r = (-1)^r(-m)_r$.

2.2.3 Covariance and Correlation

If two leaves are siblings in the partition tree, then their covariance is simply equation (6) for a Pólya Splitting model with univariate distribution $\mathcal{L}(\boldsymbol{\psi})$ given by Proposition 2. Therefore, the signs of covariance must be similar for any pairs of siblings. Hence, let us study the covariance of Y_i and Y_j that are not siblings. Using a similar argument as in Proposition 7, it can be shown that $\mathrm{Cov}(Y_i, Y_j)$ is proportional to the covariance at their common ancestor.

Proposition 8. For $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and marginals Y_i, Y_j with $\mathscr{P}(\{i\}) \neq \mathscr{P}(\{j\})$, then there is a common ancestor node $S \in \mathfrak{I}$ with $C_i, C_j \in \mathfrak{C}_S$ such that $i \in C_i$, $j \in C_j$, and

$$\mathrm{Cov}(Y_i, Y_j) = \frac{\gamma_i \gamma_j}{\gamma_{C_i} \gamma_{C_j}} \mathrm{Cov}(|\mathbf{Y}_{C_i}|, |\mathbf{Y}_{C_j}|), \tag{12}$$

where $\gamma_\ell = \prod_{k=1}^{K} \frac{\theta_{A_{k-1}}}{|\boldsymbol{\theta}_{A_k}|}$ with $A_k \in \mathrm{Path}_\ell$ for $\ell = \{i\}$, $\{j\}$, C_i or C_j.

A direct consequence of this result is that the sign of covariance between Y_i and Y_j depends only on the ancestor node. Therefore, as we will see in the running example, it is possible to have a covariance matrix with elements of different signs. Precisely, at the ancestor node S, the sign depends only on the value of c_S, i.e. the type of split, and the dispersion of $|\mathbf{Y}_S|$ with distribution given in Proposition 2. Hence, the Tree Pólya Splitting model allows for a richer covariance structure that of the Pólya Splitting model. Based on this result, and using Corollary 1, we can easily develop the covariance formula in terms of all the parameters involved in the paths from the root to the leaves and the common ancestor node.

Proposition 9. The covariance (12) is given by

$$\mathrm{Cov}(Y_i, Y_j) = \gamma_i \gamma_j \left[\left(\frac{|\boldsymbol{\theta}_S|}{|\boldsymbol{\theta}_S| + c_S} \right) \frac{\delta_S}{\gamma_S} \mu_2 - \mu_1^2 \right]$$

where $\delta_S = \prod_{k=1}^{K} \frac{\theta_{A_{k-1}} + c_{A_k}}{|\boldsymbol{\theta}_{A_k}| + c_{A_k}}$ with $A_k \in \mathrm{Path}_S$, γ_ℓ defined in Proposition 8, and μ_r the r-th factorial moment of $\mathcal{L}$.

Notice that the previous results in Proposition 8 and 9 can be adapted to any pair of subsums in the partition tree. Another interesting result describes how the ratio of covariances is equal to a ratio of expectations. Indeed, we directly have the following corollary.

Corollary 2. For $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, $A, B \in \mathfrak{I}$ such that $\mathscr{P}(A) \neq B$, $C_{A_1}, C_{A_2} \in \mathfrak{C}_A$, $C_B \in \mathfrak{C}_B$, and B is not equal or descendant of C_{A_1} or C_{A_2}, we have

$$\frac{\mathrm{Cov}\left(\left|\mathbf{Y}_{C_{A_1}}\right|, |\mathbf{Y}_{C_B}|\right)}{\mathrm{Cov}\left(\left|\mathbf{Y}_{C_{A_2}}\right|, |\mathbf{Y}_{C_B}|\right)} = \frac{\mathrm{E}\left[\left|\mathbf{Y}_{C_{A_1}}\right|\right]}{\mathrm{E}\left[\left|\mathbf{Y}_{C_{A_2}}\right|\right]} = \frac{\theta_{C_{A_1}}}{\theta_{C_{A_2}}}.$$

For the Pearson correlation, a similar result holds. Indeed, since the covariances are proportional to their ancestor node, the Pearson correlation is proportional to the same node. Using Corollary 1 for the standard deviations and Proposition 9 for the covariance, we have the following result.

Proposition 10. For $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, marginals Y_i, Y_j with $\mathscr{P}(\{i\}) \neq \mathscr{P}(\{j\})$, then there is a ancestor node $S \in \mathfrak{I}$ such that

$$\operatorname{Corr}(Y_i, Y_j) = \Lambda_i \Lambda_j \left[\left(\frac{|\boldsymbol{\theta}_S|}{|\boldsymbol{\theta}_S| + c_S} \right) \frac{\delta_S}{\gamma_S} \mu_2 - \mu_1^2 \right],$$

where

$$\Lambda_\ell = \sqrt{\frac{\gamma_\ell}{\delta_\ell \mu_2 + \mu_1 (1 - \gamma_\ell \mu_1)}},$$

and with γ_ℓ and δ_ℓ defined in Proposition 8 and 9 respectively.

Running example

Returning to our running example, let us calculate the covariance between Y_6 and Y_9. As presented in Figure 6, the ancestor node is given by $S = \{4, \ldots, 10\}$ with $C_6 = \{6, 7\}$ and $C_9 = \{8, 9, 10\}$. By Propositions 8 and 9, we have

$$\begin{aligned}
\gamma_S &= \delta_S = \pi_{4:10} \\
|\boldsymbol{\theta}_S| &= \theta_{4,5} + \theta_{6,7} + \theta_{8,9,10} \\
\gamma_6 &= \pi_{4:10} \cdot \frac{\theta_{6,7}}{|\boldsymbol{\theta}_S|} \cdot \frac{\theta_6}{\theta_6 + \theta_7} \\
\gamma_9 &= \pi_{4:10} \cdot \frac{\theta_{8,9,10}}{|\boldsymbol{\theta}_S|} \cdot \frac{\theta_{9,10}}{\theta_8 + \theta_{9,10}} \cdot \pi_9.
\end{aligned}$$

Using the identity $\mu_r = (\alpha)_r p^r / (1-p)^r$ for a $\mathcal{NB}(\alpha, p)$ distribution, the covariance is given by

$$\begin{aligned}
\operatorname{Cov}(Y_6, Y_9) &= \left[\frac{\theta_6}{\theta_6 + \theta_7} \right] \left[\pi_9 \frac{\theta_{9,10}}{\theta_8 + \theta_{9,10}} \right] \operatorname{Cov}(Y_4 + Y_5, Y_8 + Y_9 + Y_{10}) \\
&= \alpha \left(\frac{p}{1-p} \right)^2 \left[(\alpha + 1) \left(\frac{|\boldsymbol{\theta}_S|}{|\boldsymbol{\theta}_S| + 1} \right) - \alpha \right] \gamma_6 \gamma_9 \\
&= \alpha \left(\frac{p}{1-p} \right)^2 \left(\frac{|\boldsymbol{\theta}_S| - \alpha}{|\boldsymbol{\theta}_S| + 1} \right) \gamma_6 \gamma_9. \qquad (13)
\end{aligned}$$

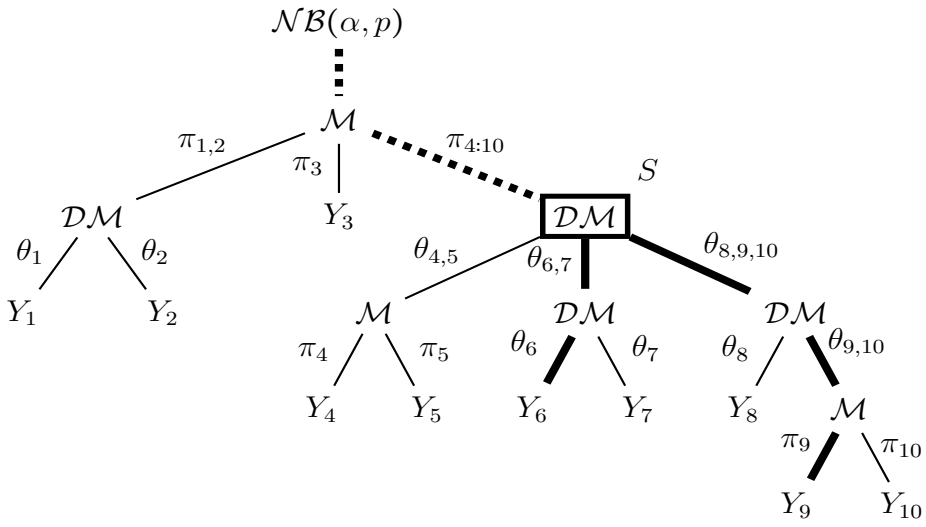

FIGURE 6 – Covariance between Y_6 and Y_9 in our running example

Notice that since $\mathcal{L} = \mathcal{NB}(\alpha, p)$, the overdispersion is preserved throughout the tree by Property 2. Therefore, it is possible to have different covariance signs. By Theorem 6 of Peyhardi et al. [2021], the distribution of $|\mathbf{Y}_S|$ is given by

$$|\mathbf{Y}_S| \sim \mathcal{NB}\left(\alpha, \frac{p\pi_{4:10}}{1 - p + p\pi_{4:10}}\right).$$

Using this marginal in equation (13), the covariance can either be negative, positive or null whether α is greater, smaller or equal to $|\boldsymbol{\theta}_S|$ respectively. Moreover, since the negative binomial is the only distribution that can induce independence for Dirichlet-multinomial Splitting, Y_6 and Y_9 are independent if and only if $\alpha = |\boldsymbol{\theta}_S|$. These dependencies are also true for any pair of random variables between the subsets $\{Y_4, Y_5\}$, $\{Y_6, Y_7\}$ and $\{Y_8, Y_9, Y_{10}\}$ since their common ancestor node is still S. In this case, a null correlation truly indicates independence.

For the sake of further investigation, suppose that $\alpha = |\boldsymbol{\theta}_S|$, but $\alpha > \theta_1 + \theta_2$. Given that $Y_1 + Y_2$ is also a negative binomial random variable, a similar argument leads us to conclude that Y_1 and Y_2 are negatively correlated, whereas Y_6 and Y_9 are independent.

For a concrete example, let us fix the following parameters.

$$\begin{aligned} \alpha &= 10 & \pi_{4:10} &= 0.6 & \theta_6 &= 0.8 \\ p &= 0.95 & \theta_{4,5} &= 3 & \theta_7 = \theta_8 &= 1 \\ \pi_{1,2} = \pi_9 &= 0.3 & \theta_{6,7} = \theta_{8,9.10} &= 3.5 & \theta_{9,10} &= 2.5 \\ \pi_3 &= 0.1 & \pi_4 = \pi_5 &= 0.5 & \pi_{10} &= 0.7 \\ \theta_1 = \theta_2 &= 1.5 \end{aligned} \tag{14}$$

Using Proposition 10, its correlation is presented in Figure 7.

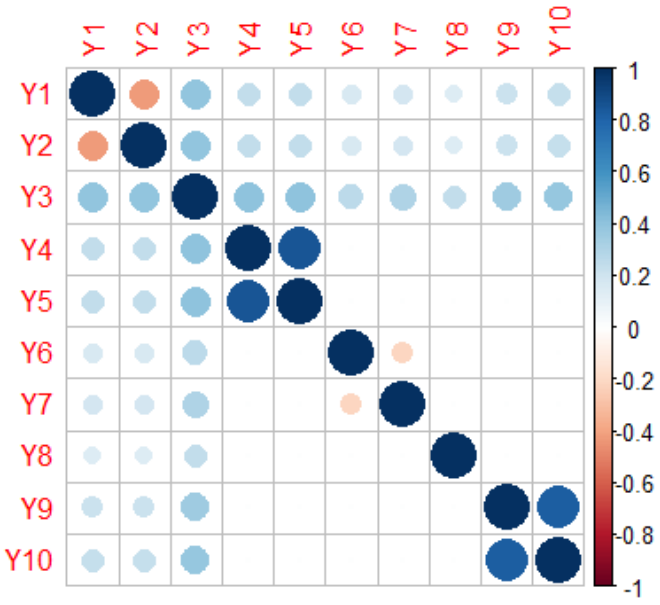

FIGURE 7 – Correlation plot of our running example with parameters (14) and $\mathcal{L} = \mathcal{NB}(\alpha, p)$.

To conclude this section, let us study the same example, but with $\mathcal{L}$ as a Dirac at m. With this simple modification the covariance at the root must be negative in this case since $\mathcal{D}_m$ is underdispersed. Furthermore, it is possible that all marginals remain underdispersed by Property 2. If the hypergeometric distribution is introduced in this example, then it

would be possible to have different signs of covariance. Finally, for comparison, suppose $m = 100$ and the tree parameters take the same values as before. Again, by Proposition 10, the correlation is presented in Figure 8 for the Dirac at m.

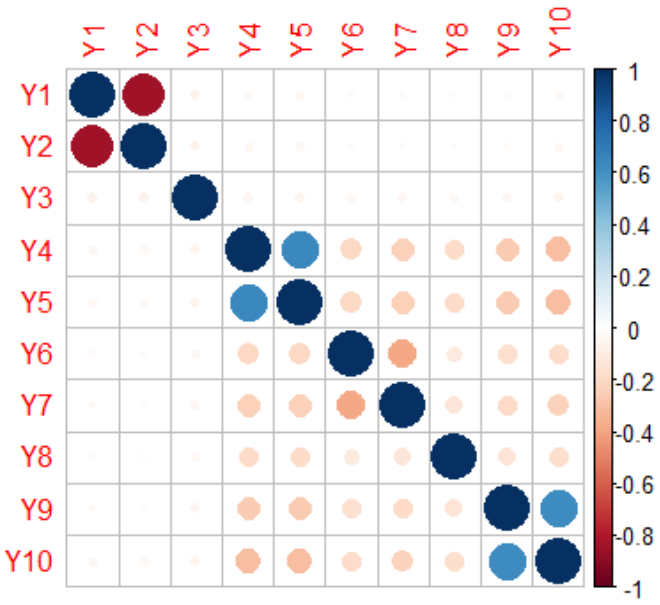

FIGURE 8 – Correlation plot of our running example with parameters (14) and $\mathcal{L} = \mathcal{D}_{100}$.

2.2.4 Log-likelihood decomposition

A direct consequence of Definition 4 is the decomposition of the log-likelihood with respect to the partition tree. Indeed, if all the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are unrelated, the log-likelihood of (7) is given by

$$\log\left[\mathbf{p}(|\mathbf{Y}| = n)\right] + \sum_{A \in \mathfrak{I}} \left(\sum_{C \in \mathfrak{C}_A} \log(\theta_C)_{(n_C, c_A)} - \log(|\boldsymbol{\theta}_A|)_{(n_A, c_A)} \right) + \text{constant}. \tag{15}$$

Therefore, the maximum likelihood estimators (MLE), if they exist, of the whole Tree

Pólya Splitting can be obtained by combining the MLE of $\boldsymbol{\psi}$ based on $|\mathbf{Y}| \sim \mathcal{L}(\boldsymbol{\psi})$ and each Pólya on $\mathfrak{T}$ separately. Such a decomposition facilitates the estimation, but also model selection. Indeed, both the AIC and BIC scores of the whole model are simply the sum of those at each node. This divide-and-conquer approach greatly simplifies the inference.

3 Analysis of a Trichoptera data set

In this section, we fit a Tree Pólya Splitting distribution with a fixed partition tree $\mathfrak{T}$ to the Trichoptera data set provided by Usseglio-Polatera and Auda [1987], comparing it to another Splitting model with a multinomial, Dirichlet-multinomial and generalized Dirichlet-multinomial tree structures, as well as to the multivariate Poisson-lognormal distribution [Aitchison and Ho, 1989; Chiquet et al., 2021]. Additionally, we fit a Tree Pólya Splitting distribution where $\mathfrak{T}$ is constructed using the data. The Trichoptera data set consists of $J = 17$ species' abundances and 9 covariates collected from $n = 49$ sites between 1959 and 1960. For sake of simplicity, no covariates are used in this application. However, it's important to note that incorporating them into the model is feasible. For the total distribution, the data exhibits overdispersion. Indeed, the empirical mean and variance of the sums are respectively 158.73 and 226617.50. Likewise, all species except three exhibit empirical overdispersion. Specifically, the species *Che*, *Hyc*, and *Hys* appear to be underdispersed.

As previously explained, the log-likelihood of the Tree Pólya can be decomposed with respect to the partition tree. Therefore, the parameters can be estimated step-by-step starting by the distribution of the total $\mathcal{L}$, and then each Pólya Splitting in the tree. Several distributions are at our disposal for $\mathcal{L}$. In particular, since the total appears to be overdispersed, one can consider a Poisson mixture, i.e. $|\mathbf{Y}|$ given λ is a Poisson distribution

with mean λ [e.g. Karlis and Xekalaki, 2005]. For the data, we fix $\mathcal{L}$ to be a negative binomial and we estimate its parameters by MLE using the `R` package `MASS` [Ripley et al., 2013]. This yields a negative binomial with parameters $\alpha = 0.478$, $p = 0.997$ and an AIC of 575.016. Since this distribution has an unbounded support, either a multinomial or Dirichlet-multinomial can be adjusted at each internal node of $\mathfrak{T}$ using the `R` package `MGLM` [Zhang et al., 2017].

Now, only the partition tree $\mathfrak{T}$ remains to be fixed. A natural approach is to use evolutionary information concerning the Trichoptera. Indeed, since the data consists of $J = 17$ different species, it is possible to regroup them into respective families. Precisely, we can divide the 17 species into 5 different groups based on the information provided by Usseglio-Polatera and Auda [1987]. With this structure, we adjust either a multinomial or a Dirichlet-multinomial at each node based on the AIC. However, should the Dirichlet-multinomial parameter estimates fail to converge, we opt for a multinomial distribution. Indeed, when $\mathcal{DM}(\boldsymbol{\theta})$ is such that each $\theta_j \to \infty$ and $\theta_j/|\boldsymbol{\theta}| \to \pi_j \in (0,1)$ for all $j \in \{1, \ldots, J\}$, then the Dirichlet-multinomial converges to a multinomial with parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$.

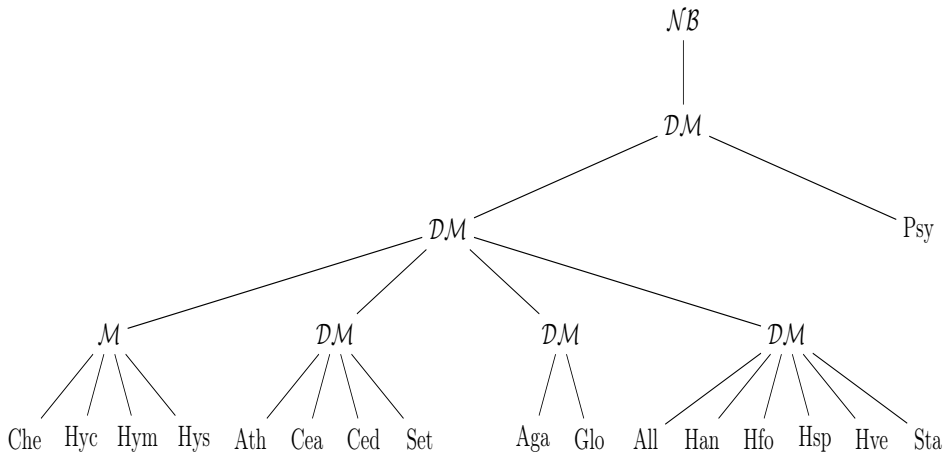

FIGURE 9 – Tree Pólya Splitting fitted to the Trichoptera data set with a fixed partition tree.

With these criteria, we have the Tree Pólya model presented in Figure 9. By decomposition of the AIC with respect to the partition tree, the adjusted Tree Pólya Splitting has a AIC of 2465.85. If we adjust the data to a multivariate Poisson-lognormal using the `R` package `PLNmodels` [Chiquet et al., 2021], we obtain an AIC of 2599.63. For this application, 170 parameters are needed for the Poisson-lognormal compare to 23 parameters for the Tree Pólya Splitting. Therefore, the proposed model is simpler and equally adequate according to the AIC score. If instead of the partition tree in Figure 9, we have used the structure of a multinomial, Dirichlet-multinomial or generalized Dirichlet-multinomial, then we would obtain an AIC of 6362.20, 2494.87, and 2460.70 respectively. Notice the latter is slightly better then the proposed tree structure. In order to find a better partition tree, we must build it using the data. Since an exhaustive search of every possible trees is numerically inconceivable, a searching algorithm must be used to efficiently find a suitable structure. Again, since the log-likelihood of the whole model is the sum of log-likelihoods at each internal node, the tree can be simply built step-by-step by summing the AIC. In the following, we present an alternative approach to construct $\mathfrak{T}$.

The search algorithm begins with a standard Dirichlet-multinomial fit. We first test if adding a child node improves the model. To do so, we create a new child node with a Dirichlet-multinomial split of two leaves and test every combination possible. If none of the combinations improve the AIC, we return to the standard Dirichlet-multinomial fit, and stop the search. Otherwise, the best combination is selected and we continue transferring one leaf from the parent node to this new node as long the AIC gets better. When these transfers stop, we test again if adding another child node from the root improves the model and transfer the leaves again. All these steps are repeated until none remain available or if the AIC measure does not improve. Finally, this process is repeated to all the new internal nodes that were previously created and the search ultimately stops as above. While this search algorithm provides a suitable tree structure, it may not yield

the optimal one. Firstly, it is possible that a tree structure exists with a better AIC score, which cannot be achieved by this algorithm. Secondly, each node in the tree has a Dirichlet-multinomial distribution. The model may be improved by changing some nodes in $\mathcal{I}$ to a multinomial. For each node, we test whether the AIC improves with a multinomial or if the parameters of some Dirichlet-multinomial diverges. Thanks to this approach, the resulting model is presented in Figure 10.

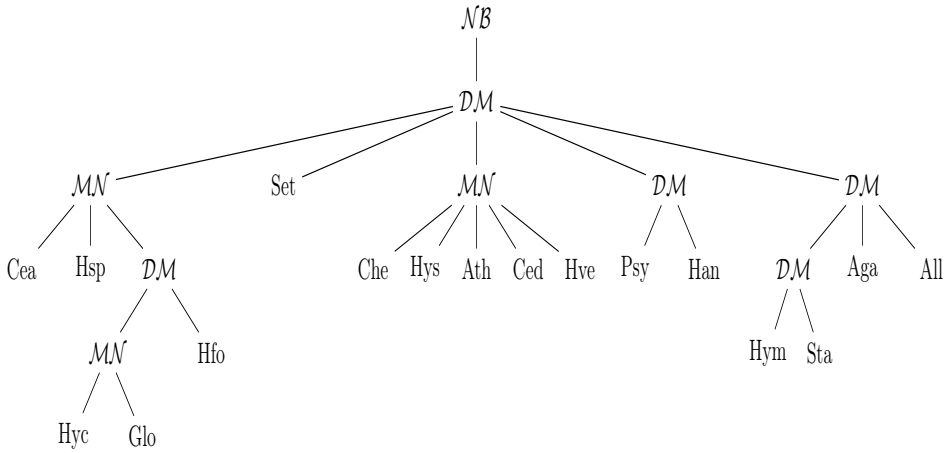

FIGURE 10 – Tree Pólya Splitting fitted to the Trichoptera data set with a partition tree search.

Using the same negative binomial distribution for $\mathcal{L}$, the AIC of this latest model is 2380.77, and still only needs 23 parameters. Each fitted model is presented in Table 2. Finally, the correlations can be easily calculated by combining Proposition 10 and the estimated parameters. See Figure 11.

Model	Nb. Parameters	AIC
Multivariate Poisson-lognormal	170	2599.63
Multinomial Splitting	18	6362.20
Dirichlet-multinomial Splitting	19	2494.87
Generalized Dirichlet-multinomial Splitting	34	2460.70
Fixed partition tree (Figure 9)	23	2465.85
Partition tree search (Figure 10)	23	2380.77

TABLE 2 – Fitted models to the Trichoptera data set

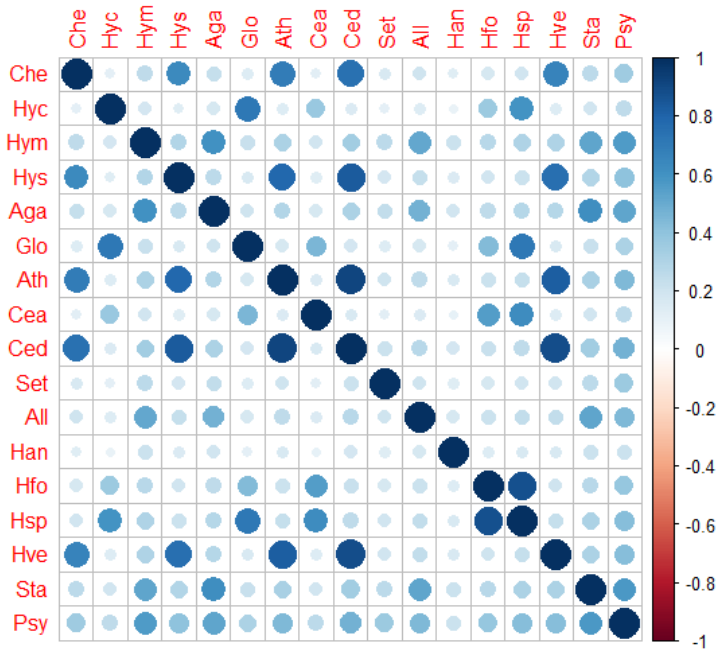

FIGURE 11 – Correlation plot of the Tree Pólya Splitting fitted to the Trichoptera data set with a partition tree search.

Discussion and perspectives

The simplicity and versatility of the Pólya Splitting model proposed by Jones and Marchand [2019] and Peyhardi et al. [2021] have been thoroughly expanded in this work.

The Tree Pólya Splitting model provides not only a generalization of the latter, but also allows for more diverse correlation structure. The partition tree of the proposed model provides a convenient model for inference, a simpler parameterization, and straightforward interpretations. This paper provides the basis of Tree Pólya Splitting models and further issues remain to be explored. Initially, Peyhardi and Fernique [2017] studied the probabilistic graphical model associated to each type of Pólya Splitting. They proved that their graphs are either complete, meaning there is an edge between all nodes, or empty, meaning no edges are present. Specifically, Peyhardi [2023] showed that the probabilistic graphical model of a Pólya Splitting distribution is empty if and only if $\mathcal{L} = \mathcal{PS}$, while being complete otherwise. Furthermore, they extended this result for a broader class of Splitting distributions, which utilize the quasi-Pólya distribution [e.g. Janardan and Schaeffer, 1977]. Given that the Tree Pólya Splitting model exhibits various dependencies, it should lead to more complex graphs than those that are complete or empty, thus bringing interesting avenues to the problem of learning graphical models with discrete variables.

Zero-inflation for multivariate count data would be interesting avenue to explore as well. Several model, including those presented by Liu and Tian [2015], Santana et al. [2022], and Zeng et al. [2023], have been proposed. Similarly, Tang and Chen [2018] provide a solution to zero-inflation for the generalized Dirichlet-multinomial model. In their work, each combination of zero-inflation is made possible thanks to the underlying binary cascade tree. Precisely, they use the zero-inflated beta-binomial distribution at each internal node, which can be interpreted as zero-inflation at all the left leaves of the tree. A generalization of this idea could be made for Tree Pólya Splitting, but instead zero-inflation could be modelled on any branch of the tree. Moudjieu et al. (in preparation) explore this particular idea for the binary Dirichlet-tree multinomial model.

Finally, analysis of extreme values for the Pólya and Tree Pólya Splittings could bring

some interesting results. Indeed, the field of extreme value theory provides a wide range of results for multivariate continuous distributions, but is lacking in terms of multivariate discrete distributions. Feidt et al. [2010] attempted to provide some answers to this problem using extreme copulas. Valiquette et al. [2023] also explored this question, but for univariate Poisson mixtures. Given that the Tree Pólya Splitting model combines a univariate discrete distribution with a tree singular distribution, integrating both their results in this model could offer valuable insights into the chanllenges of modelling multivariate discrete extreme.

Acknowledgments

This research was supported by the GAMBAS project funded by the French National Research Agency (ANR-18-CE02-0025) and the French national programme LEFE/INSU. Éric Marchand's research is supported in part by the Natural Sciences and Engineering Research Council of Canada.

Références

J. Aitchison and C. H. Ho. The multivariate Poisson-lognormal distribution. *Biometrika*, 76(4) :643–653, 1989.

X. Bry, C. Trottier, F. Mortier, and G. Cornu. Component-based regularization of a multivariate glm with a thematic partitioning of the explanatory variables. *Statistical Modelling*, 20(1) :96–119, 2020.

A. Castañer, M.M. Claramunt, C. Lefèvre, and S. Loisel. Discrete Schur-constant models. *Journal of Multivariate Analysis*, 140 :343–362, 2015.

J. Chen and H. Li. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7(1) :418 – 442, 2013.

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

R. J. Connor and J. E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325) :194–206, 1969.

R.A. Davis, K. Fokianos, S. H. Holan, H. Joe, J. Livsey, R. Lund, V. Pipiras, and N. Ravishanker. Count time series : A methodological review. *Journal of the American Statistical Association*, 116(535) :1533–1547, 2021.

S. Y. Dennis. On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Communications in Statistics - Theory and Methods*, 20(12) :4069–4081, 1991.

F. Eggenberger and G. Pólya. Über die statistik verketteter vorgänge. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 3(4) :279–289, 1923.

A. Feidt, C. Genest, and J. Nešlehová. Asymptotics of joint maxima for discontinuous random variables. *Extremes*, 13 :35–53, 2010.

K. G. Janardan and D. J. Schaeffer. A generalization of Markov-Pólya distribution its extensions and applications. *Biometrical Journal*, 19 :87–106, 1977.

H. Joe. Time series models with univariate margins in the convolution-closed infinitely divisible class. *Journal of Applied Probability*, 33(3) :664–677, 1996.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete multivariate distributions*, volume 165. Wiley New York, 1997.

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

D. Karlis and E. Xekalaki. Mixed Poisson distributions. *International Statistical Review*, 73(1) :35–58, 2005.

Y. Liu and G.L. Tian. Type I multivariate zero-inflated Poisson distribution with applications. *Computational Statistics & Data Analysis*, 83 :200–222, 2015.

F. Olver, D. Lozier, R. Boisvert, and C. Clark. *The NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY, 2010.

O. Ovaskainen and J. Soininen. Making more out of sparse data : hierarchical modeling of species communities. *Ecology*, 92(2) :289–295, 2011.

J. Peyhardi. On quasi Pólya thinning operator. *Brazilian Journal of Probability and Statistics*, 2023.

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

J. Peyhardi, P. Fernique, and J. B. Durand. Splitting models for multivariate count data. *Journal of Multivariate Analysis*, 181 :104677, 2021.

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

C. R. Rao. On discrete distributions arising out of methods of ascertainment. *Sankhyā : The Indian Journal of Statistics, Series A (1961-2002)*, 27(2/4) :311–324, 1965.

B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, and D. Firth. Package 'MASS'. *Cran r*, 538 :113–120, 2013.

R. A. Santana, K. S. Conceição, C. A. R. Diniz, and M. G. Andrade. Type I multivariate zero-inflated COM–Poisson regression model. *Biometrical Journal*, 64(3) :481–505, 2022.

M. Spivey. The Chu-Vandermonde Identity via Leibniz's Identity for Derivatives. *The College Mathematics Journal*, 47(3) :219–220, 2016.

J. Tang and A.K Gupta. On the distribution of the product of independent beta random variables. *Statistics & Probability Letters*, 2(3) :165–168, 1984.

Y. Tang, L. Ma, and D. L. Nicolae. A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data. *The Annals of Applied Statistics*, 12(1) :1 – 26, 2018.

Z.-Z. Tang and G. Chen. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4) :698–713, 06 2018.

P Usseglio-Polatera and Y Auda. Influence des facteurs météorologiques sur les résultats de piégeage lumineux. In *Annales de Limnologie-International Journal of Limnology*, volume 23, pages 65–79. EDP Sciences, 1987.

S. Valiquette, G. Toulemonde, J. Peyhardi, E. Marchand, and F. Mortier. Asymptotic tail properties of Poisson mixture distributions. *Stat*, 12(1) :e622, 2023.

T. Wang and H. Zhao. A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics*, 73(3) :792–801, 2017.

D. I. Warton, F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K.C. Hui. So many variables : Joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12) :766–779, 2015.

R. Winkelmann. *Econometric analysis of count data*. Springer Science & Business Media, 2008.

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

Y. Zeng, D. Pang, H. Zhao, and T. Wang. A zero-inflated logistic normal multinomial model for extracting microbial compositions. *Journal of the American Statistical Association*, 118(544) :2356–2369, 2023.

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

Appendix - Proofs of properties and propositions

Property 1

The proof requires the Chu-Vandermonde identity. The following proof is adapted from Spivey [2016]

Lemma 1. Let $\boldsymbol{\theta} \in \mathbb{R}^J$, $c \in \mathbb{R}$, and $n \in \mathbb{N}$. Then

$$\sum_{\mathbf{y} \in \Delta_n} \prod_{j=1}^{J} \frac{(\theta_j)_{(y_j,c)}}{y_j!} = \frac{(|\boldsymbol{\theta}|)_{(n,c)}}{n!}.$$

Proof. Since $\boldsymbol{\theta}$ and c take on real values and $(\theta)_{(n,c)} = (-c)^n (\theta)_{(n,-1)}$, it is sufficient to prove the result for $c = -1$. Using Leibniz's identity for $f_j(x) = x^{\theta_j}$, $j \in \{1, \ldots, J\}$, the n-th derivative

$$\left(\prod_{j=1}^{J} f_j(x) \right)^{(n)} = n! \sum_{\mathbf{y} \in \Delta_n} \prod_{j=1}^{J} \frac{f_j^{(y_j)}(x)}{y_j!}$$

leads to

$$\left[\frac{(|\boldsymbol{\theta}|)_{(n,-1)}}{n!} \right] x^{|\boldsymbol{\theta}|-n} = \left[\sum_{\mathbf{y} \in \Delta_n} \prod_{j=1}^{J} \frac{(\theta_j)_{(y_j,-1)}}{y_j!} \right] x^{|\boldsymbol{\theta}|-n}.$$

□

In order to find the multivariate factorial moments of $\mathbf{Y} \sim \mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, it is sufficient to find the multivariate factorial moments of the underlying Pólya, i.e. those of $\mathbf{Y}$ given $|\mathbf{Y}| = n$. Let $n \in \mathbb{N}$ and $\mathbf{r} \in \mathbb{N}^J$ such that $|\mathbf{r}| \leq n$. Then for $\mathbf{y} \geq \mathbf{r}$,

$$\begin{aligned}(-1)^{|\mathbf{r}|}(-\mathbf{y})_{\mathbf{r}} \mathbf{p}_{|\mathbf{Y}|=n}(\mathbf{y}) &= \frac{n}{(|\boldsymbol{\theta}|)_{(n,c)}} \prod_{j=1}^{J} \frac{(\theta_j)_{(y_j,c)}}{(y_j - r_j)!} \\ &= \left[\frac{n}{(|\boldsymbol{\theta}|)_{(n,c)}} \prod_{j=1}^{J} (\theta_j)_{(r_j,c)} \right] \left[\prod_{j=1}^{J} \frac{(\theta_j + c r_j)_{(y_j - r_j, c)}}{(y_j - r_j)!} \right],\end{aligned}$$

where we used the fact that $(\theta)_{(x+y,c)} = (\theta)_{(x,c)}(\theta + cx)_{(y,c)}$. Therefore, the conditional multivariate factorial moment is such that

$$\begin{aligned}(-1)^{|\mathbf{r}|}\mathrm{E}\big[(\mathbf{Y})_{\mathbf{r}} \,\big|\, |\mathbf{Y}| = n\big] &= \left[\frac{n!}{(|\boldsymbol{\theta}|)_{(n,c)}}\prod_{j=1}^{J}(\theta_j)_{(r_j,c)}\right]\sum_{\mathbf{y}\in\Delta_{n-|\mathbf{r}|}}\prod_{j=1}^{J}\frac{(\theta_j + cr_j)_{(y_j,c)}}{y_j!} \\ &= \frac{n!}{(n-|\mathbf{r}|)!}\frac{(|\boldsymbol{\theta}| + c\,|\mathbf{r}|)_{(n-|\mathbf{r}|,c)}}{(|\boldsymbol{\theta}|)_{(n,c)}}\prod_{j=1}^{J}(\theta_j)_{(r_j,c)} \\ &= \frac{(-1)^{|\mathbf{r}|}(-n)_{|\mathbf{r}|}}{(|\boldsymbol{\theta}|)_{(|\mathbf{r}|,c)}}\prod_{j=1}^{J}(\theta_j)_{(r_j,c)}\,,\end{aligned}$$

where the Chu-Vandermonde's identity has been used. We can conclude by taking the expectation of the latter equality with respect to $|\mathbf{Y}|$. □

Property 2

For this proof, we analyze the ratio

$$R := \frac{\mathrm{Var}[Y_j]}{\mathrm{E}[Y_j]} = \frac{\mu_2(\theta_j + c)}{\mu_1(|\boldsymbol{\theta}| + c)} - \frac{\mu_1\theta_j}{|\boldsymbol{\theta}|} + 1,$$

and only need to provide an inequality for the first two terms on the right-hand side. For $c = 0$, the first terms sum to

$$\frac{(\mu_2 - \mu_1^2)\theta_j}{\mu_1\,|\boldsymbol{\theta}|},$$

which is zero, positive or negative whether $\mathcal{L}$ has null, over, or under dispersion, respectively. Therefore, $R = 1$, $R > 1$ and $R < 1$ respectively. For $c = 1$, suppose $\mathcal{L}$ is overdispersed, i.e. $\mu_2 - \mu_1^2 > 0$. Then the two first terms of R are such that

$$\begin{aligned}\frac{\mu_2(\theta_j + 1)}{\mu_1(|\boldsymbol{\theta}| + 1)} - \frac{\mu_1\theta_j}{|\boldsymbol{\theta}|} &\propto \mu_2(\theta_j + 1)\,|\boldsymbol{\theta}| - \mu_1^2\theta_j(|\boldsymbol{\theta}| + 1) \\ &= (\mu_2 - \mu_1^2)\theta_j(|\boldsymbol{\theta}| + 1) + \mu_2\,|\boldsymbol{\theta}_{-j}|\,,\end{aligned}$$

which is positive. Therefore, $R > 1$. A similar argument for $c = -1$ shows that $R < 1$ when $\mu_2 - \mu_1^2 < 0$. □

Property 3

As presented, the covariance is such that

$$\mathrm{Cov}(Y_i, Y_j) = \frac{\theta_i \theta_j}{|\boldsymbol{\theta}|^2}\left[\frac{|\boldsymbol{\theta}|\,\mu_2}{|\boldsymbol{\theta}| + c} - \mu_1^2\right]. \tag{16}$$

The first two factorial moments can be expressed as follows,

$$\begin{aligned}
\mu_1 &= \frac{|\boldsymbol{\theta}|}{\theta_k}\,\mathrm{E}[Y_k], \\
\mu_2 &= \frac{|\boldsymbol{\theta}|\,(|\boldsymbol{\theta}| + c)}{\theta_k(\theta_k + c)}\,\mathrm{E}[Y_k(Y_k - 1)],
\end{aligned}$$

for any Y_k. Substituting these values in (16) yields

$$\begin{aligned}
\mathrm{Cov}(Y_i, Y_j) &= \frac{\theta_i \theta_j}{\theta_k(\theta_k + c)}\left[\mathrm{E}[Y_k(Y_k - 1)] - \left(1 + \frac{c}{\theta_k}\right)\mathrm{E}[Y_k]^2\right] \\
&= \frac{\theta_i \theta_j}{\theta_k(\theta_k + c)}\mathrm{Var}[Y_k]\left[1 - \frac{\mathrm{E}[Y_k]}{\mathrm{Var}[Y_k]}\left(1 + \frac{c}{\theta_k}\mathrm{E}[Y_k]\right)\right].
\end{aligned}$$

In particular, if $k = i$, then the covariance is simply

$$\mathrm{Cov}(Y_i, Y_j) = \frac{\theta_j}{\theta_i + c}\mathrm{Var}[Y_i]\left[1 - \frac{\mathrm{E}[Y_i]}{\mathrm{Var}[Y_i]}\left(1 + \frac{c}{\theta_i}\mathrm{E}[Y_i]\right)\right].$$

To obtain the correlation, we need to express the variance of Y_j in terms of Y_i.

$$\begin{aligned}
\mathrm{Var}[Y_j] &= \mathrm{E}[Y_j(Y_j - 1)] - \mathrm{E}[Y_j]\big(\mathrm{E}[Y_j] - 1\big) \\
&= \frac{\theta_j(\theta_j + c)}{\theta_i(\theta_i + c)}\mathrm{E}[Y_i(Y_i - 1)] - \frac{\theta_j}{\theta_i}\mathrm{E}[Y_i]\left(\frac{\theta_j}{\theta_i}\mathrm{E}[Y_i] - 1\right) \\
&= \frac{\theta_j(\theta_j + c)}{\theta_i(\theta_i + c)}\mathrm{Var}[Y_i] - \frac{\theta_j(\theta_j - \theta_i)}{\theta_i(\theta_i + c)}\mathrm{E}[Y_i]\left(1 + \frac{c}{\theta_i}\mathrm{E}[Y_i]\right) \\
&= \frac{\theta_j(\theta_j + c)}{\theta_i(\theta_i + c)}\mathrm{Var}[Y_i]\left[1 - \left(\frac{\theta_j - \theta_i}{\theta_j + c}\right)\frac{\mathrm{E}[Y_i]}{\mathrm{Var}[Y_i]}\left(1 + \frac{c}{\theta_i}\mathrm{E}[Y_i]\right)\right].
\end{aligned}$$

Finally, the results follows by letting $M = \mathrm{E}[Y_i]\big(1 + c\mathrm{E}[Y_i]/\theta_i\big)/\mathrm{Var}[Y_i]$, and since $\mathrm{Corr}(Y_i, Y_j) = \mathrm{Cov}(Y_i, Y_j)/\sqrt{\mathrm{Var}[Y_i]\mathrm{Var}[Y_j]}$. □

Property 4

It is sufficient to show this property when the parameters of $\mathbf{Y} \sim \mathcal{DM}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ allow the covariance to be positive. Without loss of generality, suppose $\theta_j \geq \theta_i$. From the proof of Property 3, we have that $1 - M(\theta_j - \theta_i)/(\theta_j + 1) \in (0, 1]$. Moreover, by hypothesis, $(\theta_j - \theta_i)/(\theta_j + 1) < 1$. Therefore, we have

$$\frac{1 - M}{\sqrt{1 - \left(\frac{\theta_j - \theta_i}{\theta_j + 1}\right) M}} \leq \frac{1 - M}{1 - \left(\frac{\theta_j - \theta_i}{\theta_j + 1}\right) M} < 1.$$

□

Propositions 1, 2 and 3

By definition, a Tree Pólya model has Pólya Splitting at each internal node. In particular, for the root Ω, each marginal is a subsum $\left|\mathbf{Y}_{C_j}\right|$, with $C_j \in \mathfrak{C}(\Omega)$, such that their distribution is given by (4) with

$$\left|\mathbf{Y}_{C_j}\right| \sim \mathcal{P}_{\Delta_n}^{[c_\Omega]}\left(\theta_{C_j}, \left|\boldsymbol{\theta}_{-C_j}(\Omega)\right|\right) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi}).$$

For each $j \in \{1, \ldots, J_\Omega\}$, we have an induced univariate distribution that defines new roots in the tree. Iterating this process, we conclude that the three propositions follow. □

Proposition 4

From Theorem 6 of Peyhardi et al. [2021], $\mathcal{B}_n(\pi) \underset{n}{\wedge} \mathcal{NB}(\alpha, p) = \mathcal{NB}\left(\alpha, \frac{p\pi}{1-p(1-\pi)}\right)$. By iterating this composition, we can easily show that

$$\bigwedge_{k=1}^{K} \mathcal{B}_{n_k}(\pi_k) \underset{n_K}{\wedge} \mathcal{NB}(\alpha, p) = \mathcal{NB}\left(\alpha, \frac{p \prod_{k=1}^{K} \pi_k}{1 - p\left(1 - \prod_{k=1}^{K} \pi_k\right)}\right).$$

Since M of those π_k are beta distributed, we have for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$ that

$$\bigwedge_{k=1}^{K} \mathcal{P}_{n_k}^{[c_k]}(\theta_k, \gamma_k) \underset{n_K}{\wedge} \mathcal{NB}(\alpha, p) = \mathcal{NB}\left(\alpha, \frac{p\gamma\pi}{1 - p(1 - \gamma\pi)}\right) \underset{\pi}{\wedge} \mathcal{PB}(\boldsymbol{\alpha}, \boldsymbol{\beta}),$$

where $\gamma = \prod_{k=1}^{K-M} \pi_k$, $\pi = \prod_{k=1}^{M} \pi_k$ and $\pi \sim \mathcal{PB}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the product beta distribution. Noticing that

$$\frac{p\gamma\pi}{1 - p(1 - \gamma\pi)} = \frac{\frac{p\gamma}{1-p(1-\gamma)}\pi}{1 - \frac{p\gamma}{1-p(1-\gamma)}(1 - \pi)},$$

we have again that

$$\begin{aligned}
\mathcal{NB}\left(r, \frac{p\gamma\pi}{1 - p(1 - \gamma\pi)}\right) \underset{\pi}{\wedge} \mathcal{PB}(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \left[\mathcal{B}_{n_M}(\pi) \underset{n_M}{\wedge} \mathcal{NB}\left(r, \frac{p\gamma}{1 - p(1-\gamma)}\right)\right] \underset{\pi}{\wedge} \mathcal{PB}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \left[\bigwedge_{k=1}^{M} \mathcal{B}_{n_k}(\pi_k) \underset{\pi}{\wedge} \mathcal{PB}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right] \underset{n_M}{\wedge} \mathcal{NB}\left(r, \frac{p\gamma}{1 - p(1-\gamma)}\right) \\
&= \bigwedge_{k=1}^{M} \mathcal{BB}_{n_k}(\alpha_k, \beta_k) \underset{n_M}{\wedge} \mathcal{NB}\left(r, \frac{p\gamma}{1 - p(1-\gamma)}\right).
\end{aligned}$$

□

Proposition 5

By the argument presented in the proof of Proposition 4, the composition of K beta-binomial distributions with a negative binomial is the same as a negative binomial composed with the product of beta random variables $\pi \sim \mathcal{PB}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. As presented in Tang and Gupta [1984], π has density

$$f_\pi(x) = (\boldsymbol{\alpha})_{\boldsymbol{\beta}} \sum_{i=0}^{\infty} \rho_i^{(K)} x^{\alpha_K - 1}(1 - x)^{|\boldsymbol{\beta}| + i - 1}; \; x \in (0, 1).$$

Since $X \sim \mathcal{NB}\left(r, \frac{p\pi}{1-p(1-\pi)}\right) \underset{\pi}{\wedge} \mathcal{PB}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, then the p.m.f. is given by

$$\begin{aligned}\mathbf{p}(n) &= (\boldsymbol{\alpha})_{\boldsymbol{\beta}}(1-p)^r \frac{(r)_n p^n}{n!} \sum_{i=0}^{\infty} \rho_i^{(K)} \int_0^1 \frac{t^{\alpha_K+n-1}(1-t)^{|\boldsymbol{\beta}|+i-1}}{(1-p(1-t))^{r+n}} dt \\ &= (\boldsymbol{\alpha})_{\boldsymbol{\beta}}(1-p)^r \frac{(r)_n p^n}{n!} \sum_{i=0}^{\infty} \rho_i^{(K)} \mathrm{B}\big(|\boldsymbol{\beta}|+i, \alpha_K+n\big) {}_2F_1\left[\begin{matrix} r+n,\ |\boldsymbol{\beta}|+i \\ \alpha_K+|\boldsymbol{\beta}|+n+i \end{matrix}; p\right],\end{aligned}$$

where we used the integral representation of ${}_2F_1\left[\begin{smallmatrix} a,\ b \\ c \end{smallmatrix}; p\right]$. For the case where $p \in (0, 1/2)$, we need the following lemma.

Lemma 2. For any $z \in \mathbb{R}$, $\pi \in (0,1)$ and $K \in \mathbb{N}_+$,

$$\sum_{i=0}^{n} \binom{n}{i} \pi^i (1-\pi)^{n-i} {}_{K+1}F_K\left[\begin{matrix} -i,\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha}+\boldsymbol{\beta} \end{matrix}; z\right] = {}_{K+1}F_K\left[\begin{matrix} -n,\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha}+\boldsymbol{\beta} \end{matrix}; \pi z\right].$$

Proof. Let $K = 1$, then

$$\begin{aligned}\sum_{i=0}^{n} \binom{n}{i} \pi^i (1-\pi)^{n-i} {}_2F_1\left[\begin{matrix} -i,\ \alpha_1 \\ \alpha_1+\beta_1 \end{matrix}; z\right] &= \int_0^1 \frac{t^{\alpha_1-1}(1-t)^{\beta_1-1}}{\mathrm{B}(\alpha_1,\beta_1)} \sum_{i=0}^{n} \binom{n}{i} (\pi(1-zt))^i (1-\pi)^{n-i} dt \\ &= \frac{1}{\mathrm{B}(\alpha_1,\beta_1)} \int_0^1 \frac{t^{\alpha_1-1}(1-t)^{\beta_1-1}}{(1-t\pi z)^{-n}} dt \\ &= {}_2F_1\left[\begin{matrix} -n,\ \alpha_1 \\ \alpha_1+\beta_1 \end{matrix}; \pi z\right].\end{aligned}$$

For $K \geq 1$, the result follows from the identity [e.g. Olver et al., 2010]

$${}_{K+2}F_{K+1}\left[\begin{matrix} -i,\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha}+\boldsymbol{\beta} \end{matrix}; z\right] = \int_0^1 \frac{t^{\alpha_1-1}(1-t)^{\beta_1-1}}{\mathrm{B}(\alpha_1,\beta_1)} {}_{K+1}F_K\left[\begin{matrix} -i,\ \boldsymbol{\alpha}_{-1} \\ \boldsymbol{\alpha}_{-1}+\boldsymbol{\beta}_{-1} \end{matrix}; zt\right] dt.$$

□

First, let us calculate the probability generating function, denoted by $G(z) := \mathbb{E}[z^X]$, of

$$\bigwedge_{k=1}^{K} \mathcal{BB}_{n_k}(\alpha_k, \beta_k). \tag{17}$$

For $K = 1$, it is well known that $G(z) = {}_2F_1\left[\begin{smallmatrix} -n_1,\ \alpha_1 \\ \alpha_1+\beta_1 \end{smallmatrix}; 1-z\right]$. Suppose for $K \geq 1$ that the generating function of (17) is given by

$$G(z) = {}_{K+1}F_K\left[\begin{matrix} -n_K,\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha}+\boldsymbol{\beta} \end{matrix}; 1-z\right]. \tag{18}$$

Then for $K+1$ and by conditioning on the last K terms of (17), we have by hypothesis

$$G(z) = \mathbb{E}\left[{}_{K+1}F_K\left[\begin{matrix} -n_K,\ \boldsymbol{\alpha}_{-(K+1)} \\ \boldsymbol{\alpha}_{-(K+1)} + \boldsymbol{\beta}_{-(K+1)} \end{matrix}; 1-z\right]\right],$$

where the expectation is taken on the last beta-binomial $\mathcal{BB}_{n_{K+1}}(\alpha_{K+1}, \beta_{K+1})$. Since the beta-binomial is a binomial mixture, the use of Lemma 2 leads to

$$\begin{aligned} G(z) &= \mathbb{E}\left[{}_{K+1}F_K\left[\begin{matrix} -n_{K+1},\ \boldsymbol{\alpha}_{-(K+1)} \\ \boldsymbol{\alpha}_{-(K+1)} + \boldsymbol{\beta}_{-(K+1)} \end{matrix}; (1-z)\pi\right]\right] \\ &= \int_0^1 \frac{\pi^{\alpha_{K+1}-1}(1-\pi)^{\beta_{K+1}-1}}{\mathrm{B}(\alpha_{K+1}, \beta_{K+1})} {}_{K+1}F_K\left[\begin{matrix} -n_{K+1},\ \boldsymbol{\alpha}_{-(K+1)} \\ \boldsymbol{\alpha}_{-(K+1)} + \boldsymbol{\beta}_{-(K+1)} \end{matrix}; (1-z)\pi\right] d\pi \\ &= {}_{K+2}F_{K+1}\left[\begin{matrix} -n_{K+1},\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha} + \boldsymbol{\beta} \end{matrix}; 1-z\right]. \end{aligned}$$

Notice that (18) exists for all $z \in \mathbb{R}$. In order to find the probability generating function of the full distribution, we only need to take the expectation of (18) with respect to $n_K \sim \mathcal{NB}(\alpha, p)$. In fact, we prove that for $z \in (2 - p^{-1}, p^{-1})$, the probability generating function is given by

$$G(z) = {}_{K+1}F_K\left[\begin{matrix} r,\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha} + \boldsymbol{\beta} \end{matrix}; \frac{p}{p-1}(1-z)\right].$$

For $K = 1$, Jones and Marchand [2019] proved this result. Suppose it is true for $K \geq 1$. Then for $K+1$, the generating function is given by

$$\begin{aligned} G(z) &= (1-p)^r \sum_{n=0}^{\infty} \frac{(r)_n p^n}{n!} {}_{K+2}F_{K+1}\left[\begin{matrix} -n,\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha} + \boldsymbol{\beta} \end{matrix}; 1-z\right] \\ &= \int_0^1 \frac{t^{\alpha_{K+1}-1}(1-t)^{\beta_{K+1}-1}}{\mathrm{B}(\alpha_{K+1}, \beta_{K+1})} {}_{K+1}F_K\left[\begin{matrix} r,\ \boldsymbol{\alpha}_{-(K+1)} \\ \boldsymbol{\alpha}_{-(K+1)} + \boldsymbol{\beta}_{-(K+1)} \end{matrix}; \frac{p}{p-1}(1-z)t\right] dt \\ &= {}_{K+2}F_{K+1}\left[\begin{matrix} r,\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha} + \boldsymbol{\beta} \end{matrix}; \frac{p}{p-1}(1-z)\right], \end{aligned}$$

which proves the result since the latter equivalence requires $\left|\frac{p}{p-1}(1-z)\right| < 1$. Moreover, to obtain the p.m.f., it suffices to evaluate the term $G^{(k)}(0)/k!$ for $p < 1/2$. □

Proposition 6

Using a similar argument as in Proposition 4, it can be shown that we can interchange the order of composition, i.e.

$$\bigwedge_{k=1}^{K} \mathcal{P}_{n_k}^{[c_k]}(\theta_k, \tau_k) \underset{n_K}{\wedge} \mathcal{D}_m = \left[\bigwedge_{k=1}^{M} \mathcal{BB}_{n_k}(\alpha_k, \beta_k) \right] \underset{n_M}{\wedge} \mathcal{B}_m(\gamma).$$

Combining Equation (18) and Lemma 2, the probability generating function of X is given by

$$G(z) = \mathrm{E}\left[\mathrm{E}\left[z^X \mid n_M\right]\right] = \mathrm{E}\left[{}_{M+1}F_M\left[\begin{matrix} -n_M, \boldsymbol{\alpha} \\ \boldsymbol{\alpha}+\boldsymbol{\beta} \end{matrix}; 1-z \right]\right] = {}_{M+1}F_M\left[\begin{matrix} -m, \boldsymbol{\alpha} \\ \boldsymbol{\alpha}+\boldsymbol{\beta} \end{matrix}; \gamma(1-z) \right],$$

for any $z \in \mathbb{R}$. Since $\mathbf{p}(n) = G^{(n)}(0)/n!$ and $\frac{d^n}{dz^n} {}_pF_q\left[\begin{matrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{matrix}; z\right] = \frac{(\boldsymbol{\alpha})_n}{(\boldsymbol{\beta})_n} {}_pF_q\left[\begin{matrix}\boldsymbol{\alpha}+n\mathbf{1}\\ \boldsymbol{\beta}+n\mathbf{1}\end{matrix}; z\right]$, we can conclude. □

Proposition 7

For each $\{j\} \in \mathfrak{L}$, there is a $\mathrm{Path}_{\{j\}}$ of length $K_{\{j\}}$. With each path, we can identify the leaves with the greatest path length. Let us note that there are at least two leaves with maximum length due to the Splitting model structure. Furthermore, those leaves can be regrouped as siblings. Without loss of generality, suppose the $m \geq 2$ first leaves have maximum length, and are siblings with a common parent node $A = \{1, \ldots, m\}$. From this set, we can use the law of total expectation yielding

$$(-1)^{|\mathbf{r}|}\mathrm{E}[(-\mathbf{Y})_{\mathbf{r}}] = (-1)^{|\mathbf{r}|}\mathrm{E}\left[(\mathbf{Y}_{-A})_{\mathbf{r}_{-A}} \mathrm{E}\left[(\mathbf{Y}_A)_{\mathbf{r}_A} \middle| \mathbf{Y}_{-A}, |\mathbf{Y}_A| \right] \right].$$

Since $\mathbf{Y}_A$ is conditionally independent of $\mathbf{Y}_{-A}$, and the distribution of $\mathbf{Y}_A$ given $|\mathbf{Y}_A|$ is Pólya with parameter $\boldsymbol{\theta}_A$, then by Property 1

$$(-1)^{|\mathbf{r}|}\mathrm{E}[(-\mathbf{Y})_{\mathbf{r}}] = \frac{\prod_{C \in \mathfrak{C}_A} (\theta_C)_{(|\mathbf{r}_C|, c_A)}}{(|\boldsymbol{\theta}_A|)_{(|\mathbf{r}_A|, c_A)}} \mathrm{E}\left[(-|\mathbf{Y}_A|)_{|\mathbf{r}_A|} (\mathbf{Y}_{-A})_{\mathbf{r}_{-A}} \right].$$

From this point, the factorial moment of the full Tree Pólya can be obtained by calculating the factorial moment of a new Tree Pólya Splitting model where the m first leaves are replaced by the leaf A, and calculating its $|\mathbf{r}_A|$-th factorial moment. By iterating this process for the next set of siblings with maximal path length, we get the product over all internal nodes as mentioned. Once this process arrives at the root Ω, the factorial moment is given by

$$(-1)^{|\mathbf{r}|}\mathrm{E}[(-\mathbf{Y})_{\mathbf{r}}] = (-1)^{|\mathbf{r}|}\mathrm{E}[(-\mathbf{Y})_{|\mathbf{r}|}] \prod_{A\in\mathcal{I}} \frac{\prod_{C\in\mathfrak{C}_A}(\theta_C)_{(|\mathbf{r}_C|,c_A)}}{(|\boldsymbol{\theta}_A|)_{(|\mathbf{r}_A|,c_A)}},$$

and the right-hand side expectation is simply $\mu_{|\mathbf{r}|}$. □

Proposition 8

By hypothesis, at least one path from a leaf to the ancestor node has a strictly positive length. Otherwise, both Y_i and Y_j are siblings. Without loss of generality, let us suppose that $\mathrm{Path}_{\{i\}}^{C_i}$ has length $K > 0$. Then, for $A_k \in \mathrm{Path}_{\{i\}}^{C_i}$, we have by Corollary 1 that

$$\mathrm{E}[Y_i] = \left(\prod_{k=1}^{K} \frac{\theta_{A_{k-1}}}{|\boldsymbol{\theta}_{A_k}|}\right) \mathrm{E}\big[\,|\mathbf{Y}_{C_i}|\,\big].$$

Secondly, by a similar argument from the previous proof, we have

$$\mathrm{E}[Y_iY_j] = \left(\prod_{k=1}^{K} \frac{\theta_{A_{k-1}}}{|\boldsymbol{\theta}_{A_k}|}\right) \mathrm{E}\big[\,|\mathbf{Y}_{C_i}|\,Y_j\big].$$

By definition, the covariance is given by

$$\mathrm{Cov}(Y_i,Y_j) = \left(\prod_{k=1}^{K} \frac{\theta_{A_{k-1}}}{|\boldsymbol{\theta}_{A_k}|}\right) \mathrm{Cov}(|\mathbf{Y}_{C_i}|\,,Y_j).$$

Moreover, by definition of γ_ℓ, the product on the right-hand side is such that $\prod_{k=1}^{K} \frac{\theta_{A_{k-1}}}{|\boldsymbol{\theta}_{A_k}|} = \frac{\gamma_i}{\gamma_{C_i}}$. Finally, if $\mathrm{Path}_{\{j\}}^{C_j}$ has length 0, we conclude. Otherwise, a similar argument on the expectations for $\mathrm{Path}_{\{j\}}^{C_j}$ yields the result. □

Proposition 9

The covariance of $|\mathbf{Y}_{C_i}|$ and $\left|\mathbf{Y}_{C_j}\right|$ depends only on the parameter $\boldsymbol{\theta}_S$ and the first two factorial moments of $|\mathbf{Y}_S|$, denoted for this proof by $\tilde{\mu}_1$ and $\tilde{\mu}_2$, at the ancestor node S. Using Corollary 1 and the definitions of γ_S and δ_S yields

$$\tilde{\mu}_1 = \gamma_S \mu_1, \qquad \tilde{\mu}_2 = \delta_S \gamma_S \mu_2.$$

From equation (6), the covariance is given by

$$\begin{aligned}
\operatorname{Cov}(|\mathbf{Y}_{C_i}|, \left|\mathbf{Y}_{C_j}\right|) &= \frac{\theta_{C_i}\theta_{C_j}}{|\boldsymbol{\theta}_S|^2}\left[\frac{|\boldsymbol{\theta}_S|}{(|\boldsymbol{\theta}_S|+c)}\tilde{\mu}_2 - \tilde{\mu}_1^2\right] \\
&= \frac{\theta_{C_i}\theta_{C_j}}{|\boldsymbol{\theta}_S|^2}\gamma_S^2\left[\frac{|\boldsymbol{\theta}_S|}{(|\boldsymbol{\theta}_S|+c)}\frac{\delta_S}{\gamma_S}\mu_2 - \mu_1^2\right] \\
&= \gamma_{C_i}\gamma_{C_j}\left[\frac{|\boldsymbol{\theta}_S|}{(|\boldsymbol{\theta}_S|+c)}\frac{\delta_S}{\gamma_S}\mu_2 - \mu_1^2\right].
\end{aligned}$$

We can conclude by combining this equality with Proposition 8. □

Proposition 10

We only need to calculate the variance of Y_ℓ for $\ell = i, j$. Again, by Corollary 1, the result follows since

$$\operatorname{Var}(Y_\ell) = \mathrm{E}[Y_\ell(Y_\ell - 1)] + \mathrm{E}[Y_\ell]\big(1 - \mathrm{E}[Y_\ell]\big) = \gamma_\ell\big(\delta_\ell\mu_2 + \mu_1(1 - \gamma_\ell\mu_1)\big).$$

□