Tree Pólya Splitting distributions for multivariate count data

Samuel Valiquette

In collaboration with
Jean Peyhardi, Éric Marchand, Gwladys Toulemonde, and Frédéric Mortier

Introduction

The analysis of multivariate count data is fundamental in various fields.

- Ecology: Joint Species Distribution Models;
- Genomics: RNA sequences;
- Insurance: Number of claims.

Introduction

An appropriate model should be:

- Flexible for various correlations;
- Simple for inference and interpretation;
- Taking into account overdispersion.

Introduction

Discrete Schur-constant [Castañer et al. (2015)]

Let $\mathbf{Y} := (Y_1, \ldots, Y_J) \in \mathbb{N}^J$ be a discrete random vector. It has a Schur-constant joint survival function if, for all $\mathbf{y} = (y_1, \ldots, y_J) \in \mathbb{N}^J$,

$$\mathbb{P}(\mathbf{Y} \geq \mathbf{y}) = S(|\mathbf{y}|),$$

where $S : \mathbb{N} \mapsto [0, 1]$ is a valid function, and $|\mathbf{y}| := y_1 + \cdots + y_J$.

Introduction

Under this hypothesis, Castañer et al. [2015] prove that $\mathbf{Y}$ is Schur-constant if there exists a univariate discrete distribution $\mathcal{L}(\psi)$ such that

$$|\mathbf{Y}| \sim \mathcal{L}(\psi) \text{ and } \mathbf{Y} \mid |\mathbf{Y}| = n \sim \mathcal{DM}_{\Delta_n}(\mathbf{1}),$$

where $\mathcal{DM}_{\Delta_n}(\mathbf{1})$ denotes the Dirichlet-multinomial distribution with parameters $\mathbf{1} = (1, \ldots, 1)$ and support $\Delta_n := \{\mathbf{y} \in \mathbb{N}^J : |\mathbf{y}| = n\}$, the discrete simplex.

Introduction

Birth–death process under extended neutral theory [Peyhardi et al. (2024)]

As an application to joint species distribution models, Peyhardi et al. [2024] studied the stationary distribution of a multivariate birth-death process under the ecological neutrality assumption.

Introduction

In this framework, they proved that the random vector $\mathbf{Y}$ at the stationary state satisfies

$$|\mathbf{Y}| \sim \mathcal{L}(\psi) \quad \text{and} \quad \mathbf{Y} \mid |\mathbf{Y}| = n \sim \mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}),$$

where $\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta})$ denotes the Pólya distribution with parameters $c \in \{-1, 0, 1\}$ and $\boldsymbol{\theta} \in \mathbb{R}_+^J$.

Introduction

These special cases can all be represented by the following construction proposed by Jones & Marchand [2019] and Peyhardi et al. [2021]:

1. The sum $|\mathbf{Y}| := Y_1 + \cdots + Y_J$ follows a univariate discrete distribution $\mathcal{L}(\psi)$;
2. Given $|\mathbf{Y}| = n$, $\mathbf{Y}$ follows a Pólya distribution on the discrete simplex Δ_n.

Introduction

They name this class of distributions the *Sums and Shares* or the *Pólya Splitting*. Various distributions are particular cases, such as:

- Poisson-multinomial distribution;
- Negative multinomial distribution;
- Multivariate generalized Waring distribution [Xekalaki (1986)];
- Counts of Bernoulli success strings [Ait Aoudia et al. (2016)].

Introduction

This class of distributions, however, is quite restrictive since **all pairwise correlations must have identical signs**.

In this work, we sought to address this limitation by generalizing the Pólya Splitting model.

Goals:

1. Define and build the *Tree Pólya Splitting* model;
2. Study its properties, particularly its dependence structure;
3. Apply the model to the Trichoptera dataset.

Contents

Pólya Splitting model

Pólya Splitting model

Let $c \in \{-1, 0, 1\}$, the generalized factorial is given by

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

As special cases, we have the following:

- Falling factorial: $(x)_{(n)} := (x)_{(n,-1)}$;
- Rising factorial: $(x)_n := (x)_{(n,1)}$.

Pólya Splitting model

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J) \in \mathbb{R}_+^J$, the probability mass function (p.m.f.) of the *Pólya* distribution $\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta})$ is given by

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

Pólya Splitting model

The three particular cases of $\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta})$ are given by

- $c = -1$: Multivariate hypergeometric distribution $\mathcal{H}_{\Delta_n}(\boldsymbol{\theta})$;
- $c = 0$: Multinomial distribution $\mathcal{M}_{\Delta_n}(\boldsymbol{\theta})$;
- $c = 1$: Dirichlet-multinomial distribution $\mathcal{DM}_{\Delta_n}(\boldsymbol{\theta})$.

Pólya Splitting model

Similarly, for the parameters θ, γ and $y \in \{0, \ldots, n\}$, the *univariate Pólya* has p.m.f.

$$\mathbb{P}(Y = y) = \frac{n!}{(\theta + \gamma)_{(n,c)}} \frac{(\theta)_{(y,c)}}{y!} \frac{(\gamma)_{(n-y,c)}}{(n-y)!} \mathbb{1}_{\{0,\ldots,n\}}(y).$$

We denote the latter by $\mathcal{P}_n^{[c]}(\theta, \gamma)$, and its particular cases are:

- $c = -1$: Hypergeometric distribution $\mathcal{H}_n(\theta, \gamma)$;
- $c = 0$: Binomial distribution $\mathcal{B}_n(\theta, \gamma)$;
- $c = 1$: Beta-binomial distribution $\mathcal{BB}_n(\theta, \gamma)$.

Pólya Splitting model

The random vector $\mathbf{Y} = (Y_1, \ldots, Y_J) \in \mathbb{N}^J$ has a *Pólya Splitting* distribution if

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

where $\wedge$ is the mixture operator. Its p.m.f. is given by

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

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

Pólya Splitting model - Marginal

Theorem [Peyhardi et al. (2021)]

Let $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and $\mathcal{I} \subseteq \{1, \dots, J\}$. The subvector $\mathbf{Y}_{\mathcal{I}}$ follows the distribution

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

In particular, the marginal distribution Y_j is given by

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

Note: The univariate marginal is equivalent to the thinning operator in discrete time series (see Peyhardi [2023]).

Pólya Splitting model - Covariance & correlation

Theorem [Peyhardi et al. (2021)]

Let $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and marginals Y_i, Y_j $(i \neq j)$. Then, the covariance is given by

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

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

$$\mu_r := \mathrm{E}\left[(|\mathbf{Y}|)_{(r)} \right].$$

Note: $\mu_2 - \mu_1^2 = \mathrm{Var}\left[|\mathbf{Y}|\right] - \mathrm{E}\left[|\mathbf{Y}|\right]$, which determines the dispersion of $\mathcal{L}$ and the sign of the covariance.

Pólya Splitting model - Covariance & correlation

Theorem [Valiquette et al. (2025)]

Let $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and marginals Y_i, Y_j $(i \neq j)$. Then, the Pearson correlation is given by

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

where

$$M_k = \frac{\mu_1 \left(1 + \frac{c}{|\boldsymbol{\theta}|}\mu_1\right)}{\mu_2 \left(\frac{\theta_k + c}{|\boldsymbol{\theta}| + c}\right) + \mu_1 \left(1 - \mu_1 \frac{\theta_k}{|\boldsymbol{\theta}|}\right)}, \quad k = i, j,$$

and $\mathrm{sgn}(x) = \mathbb{1}_{[0,\infty)} - \mathbb{1}_{(-\infty,0]}$.

Pólya Splitting model - Covariance & correlation

For $c = 1$, we can find an upper bound for this correlation, which generalizes the result of Jones & Marchand [2019].

Theorem [Valiquette et al. (2025)]

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

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

Tree Pólya Splitting model

Tree Pólya Splitting - Definitions

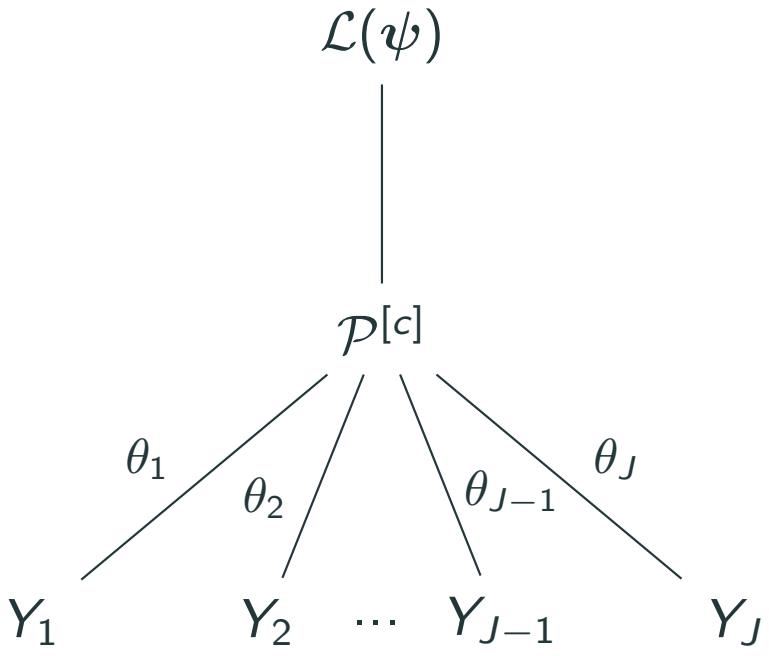
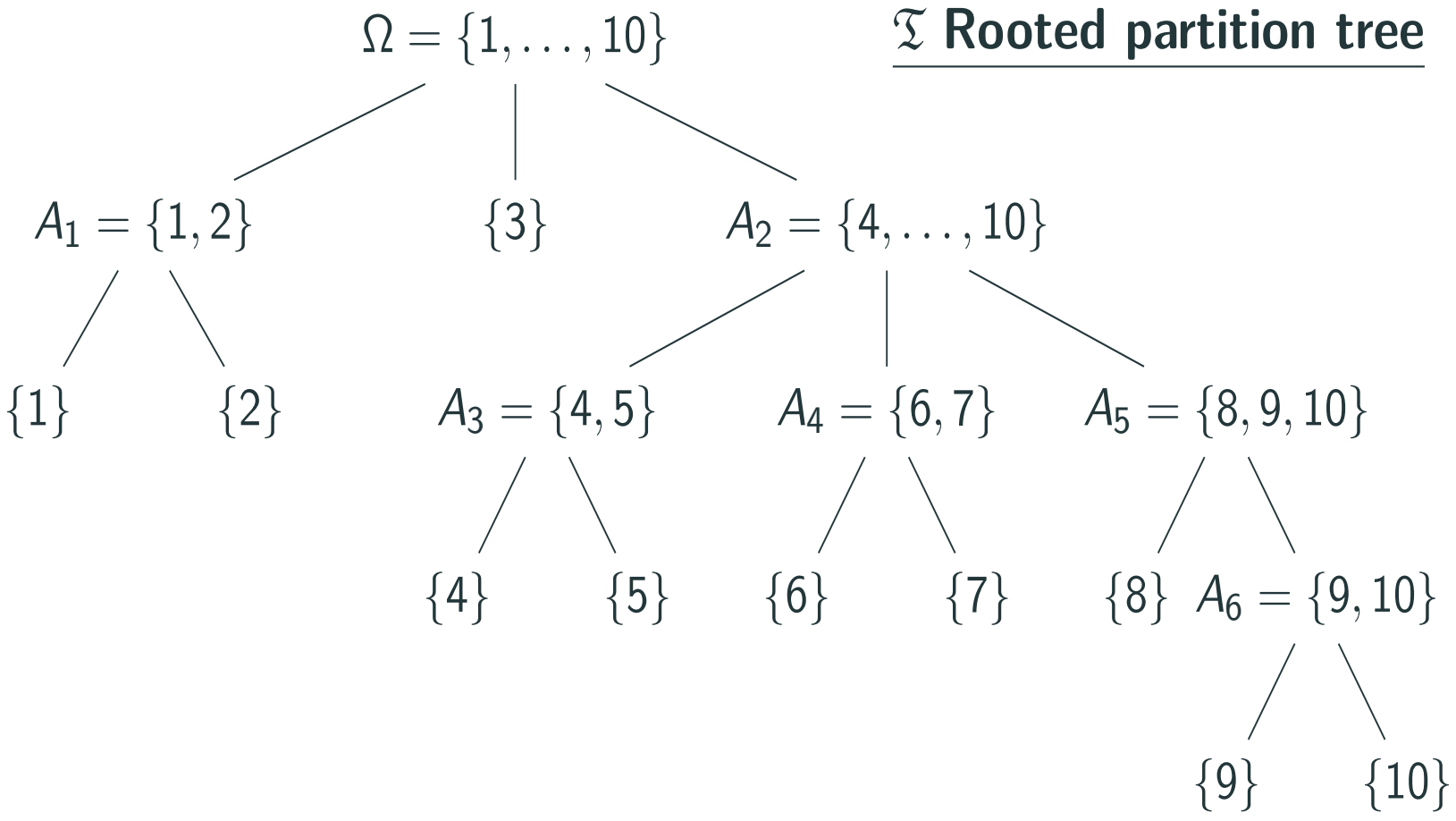
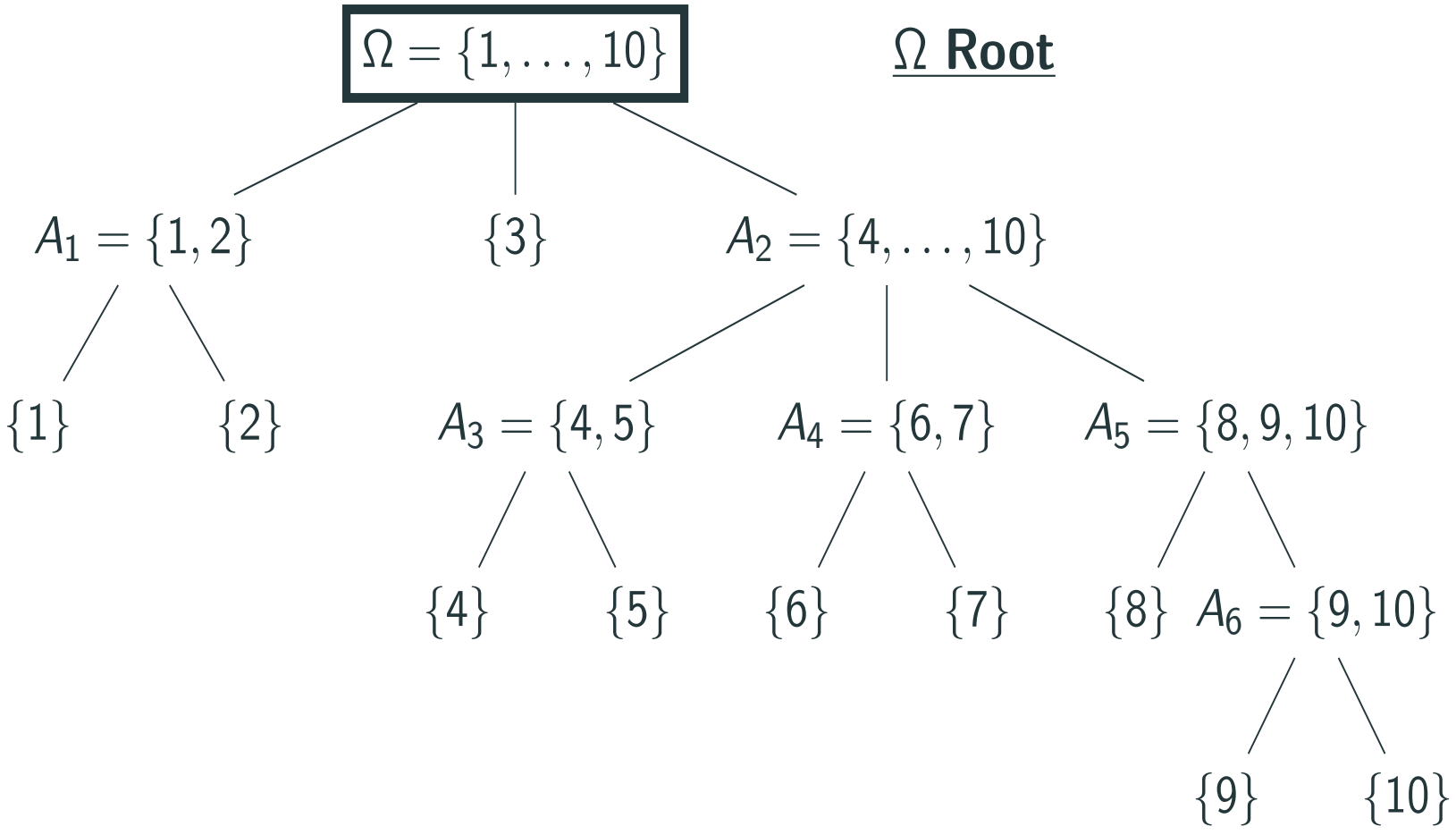


Figure 1: Representation of the Pólya Splitting model

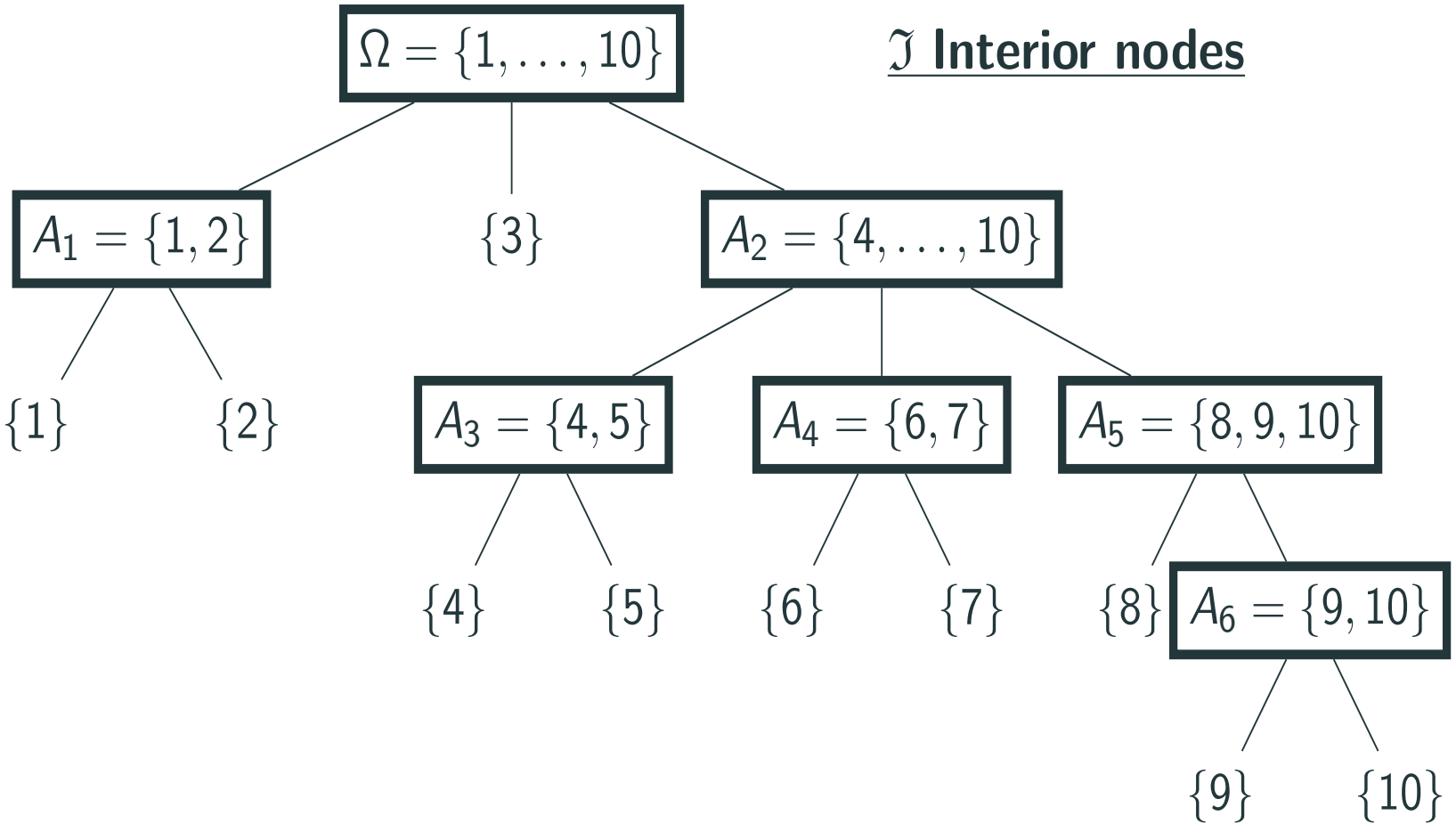
Tree Pólya Splitting - Definitions



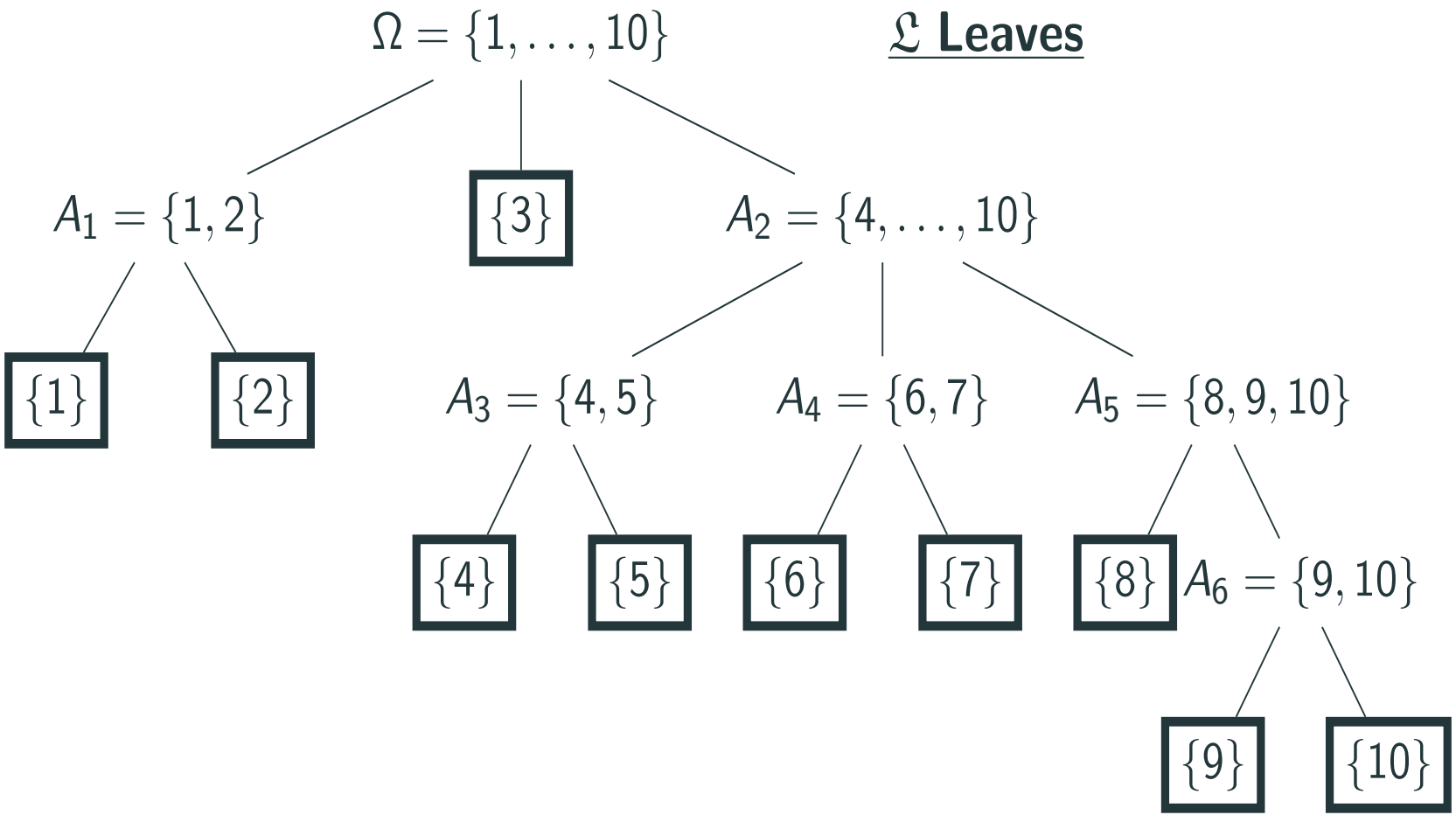
Tree Pólya Splitting - Definitions



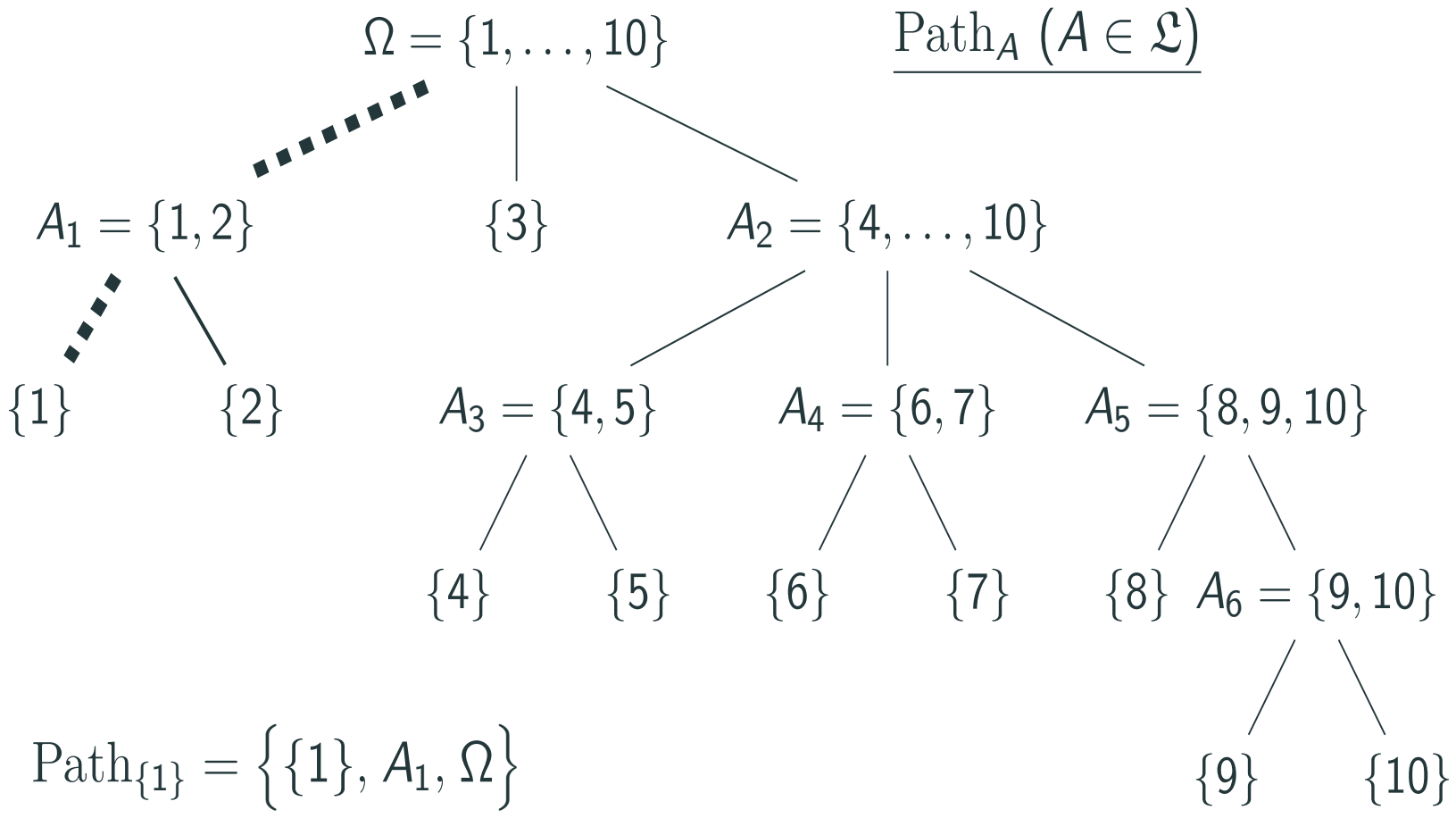
Tree Pólya Splitting - Definitions



Tree Pólya Splitting - Definitions

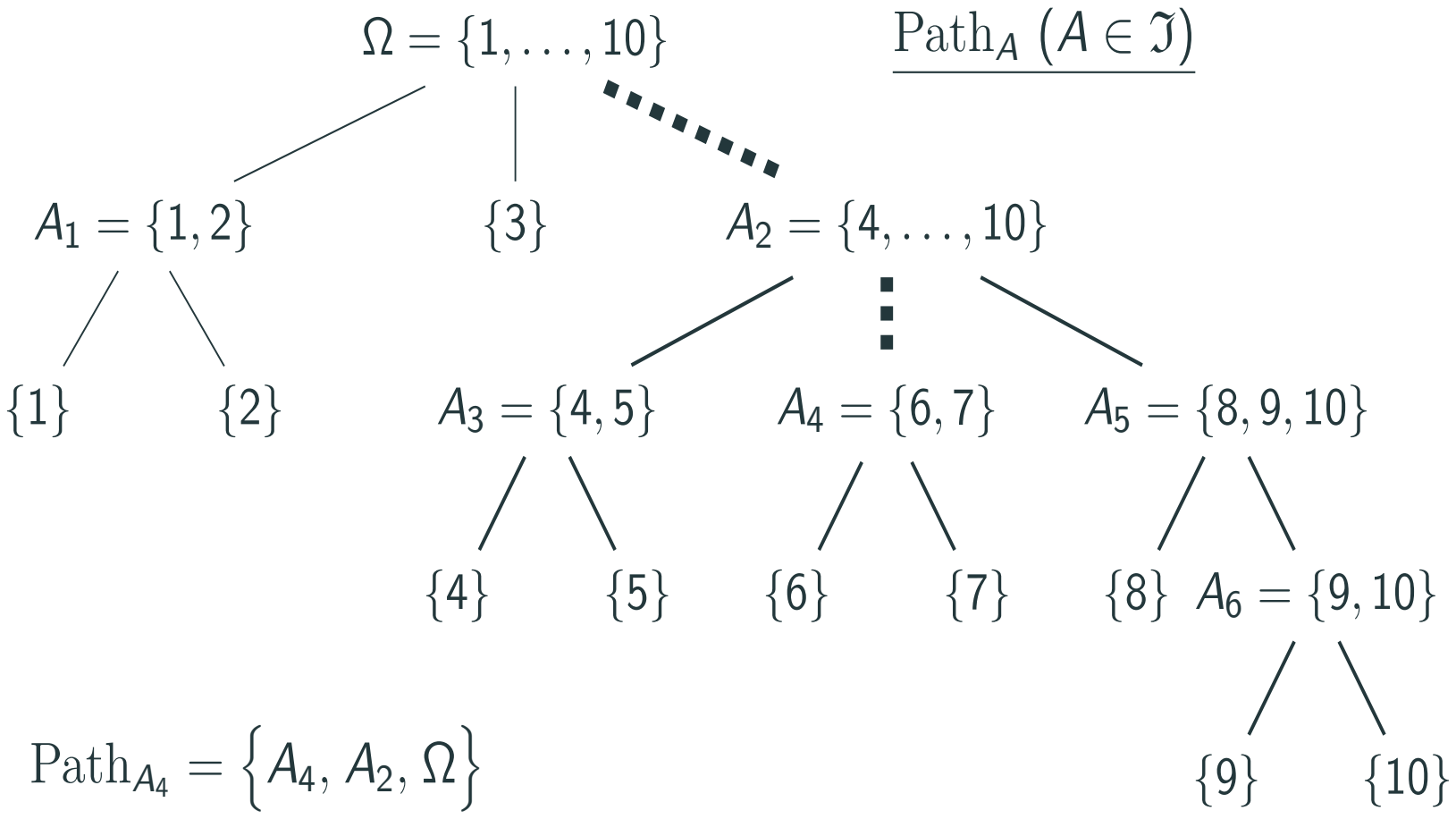


Tree Pólya Splitting - Definitions



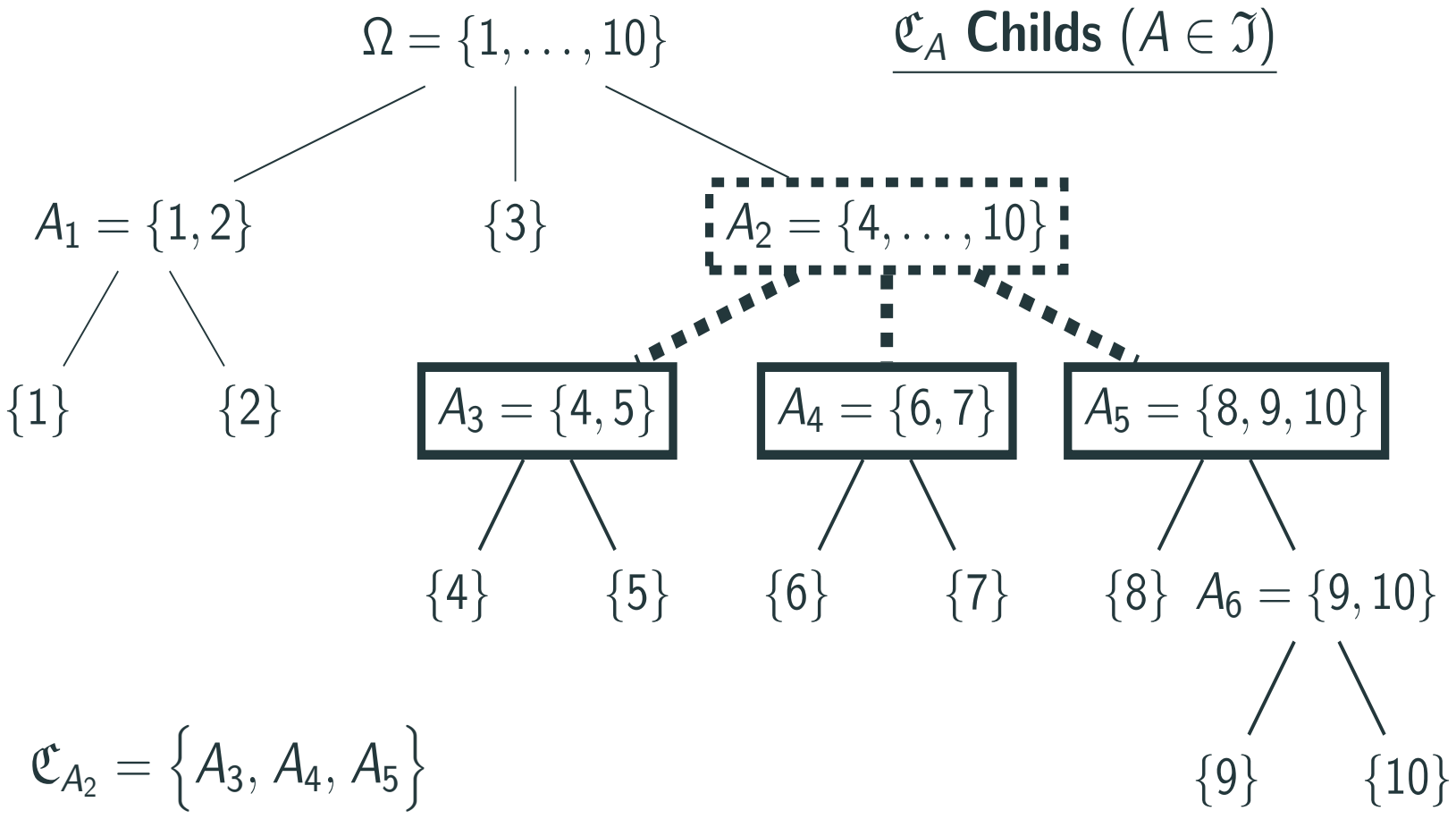
$$\mathrm{Path}_{\{1\}} = \Big\{\{1\}, A_1, \Omega\Big\}$$

Tree Pólya Splitting - Definitions



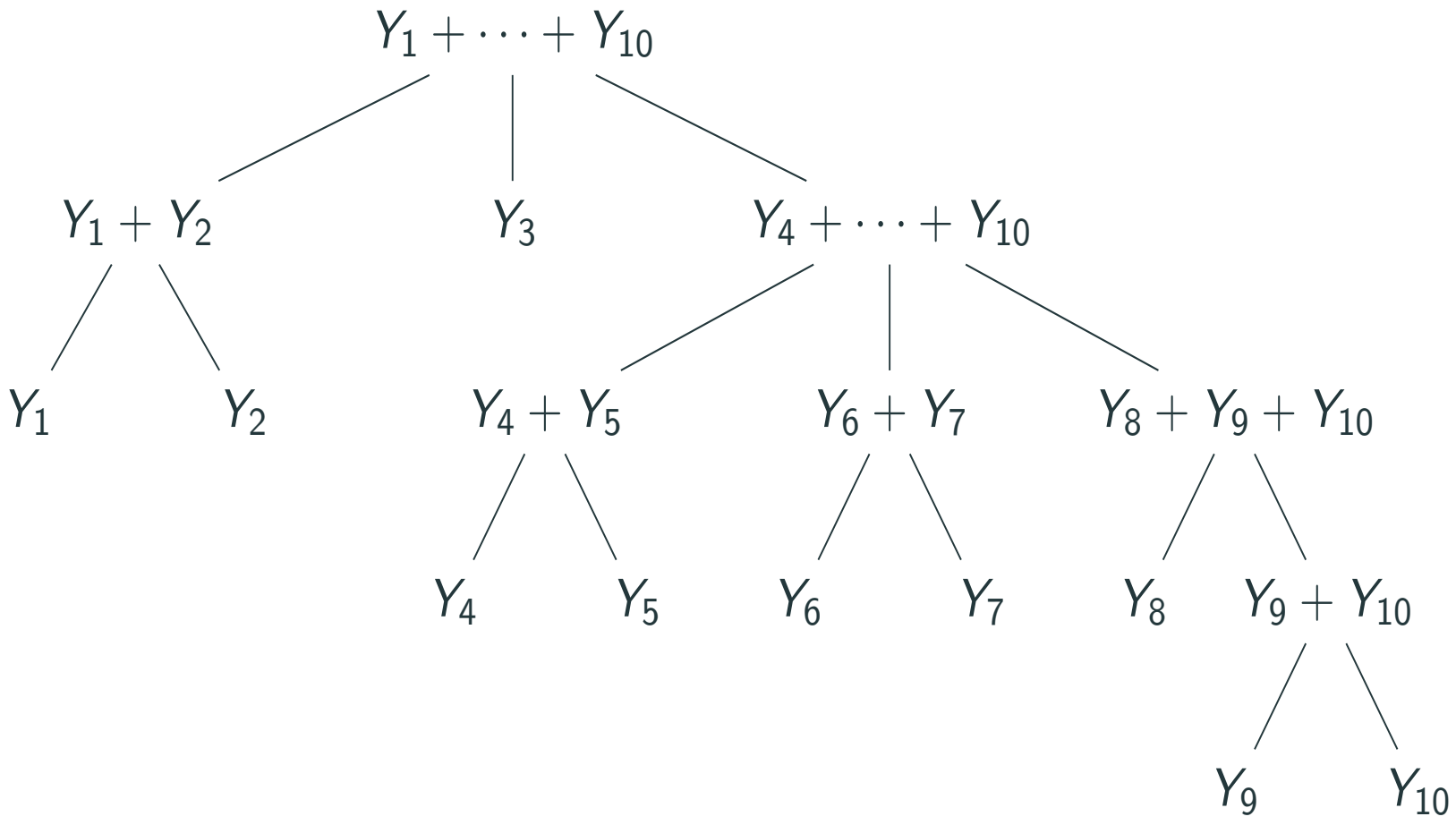
$$\mathrm{Path}_{A_4} = \left\{A_4, A_2, \Omega\right\}$$

Tree Pólya Splitting - Definitions

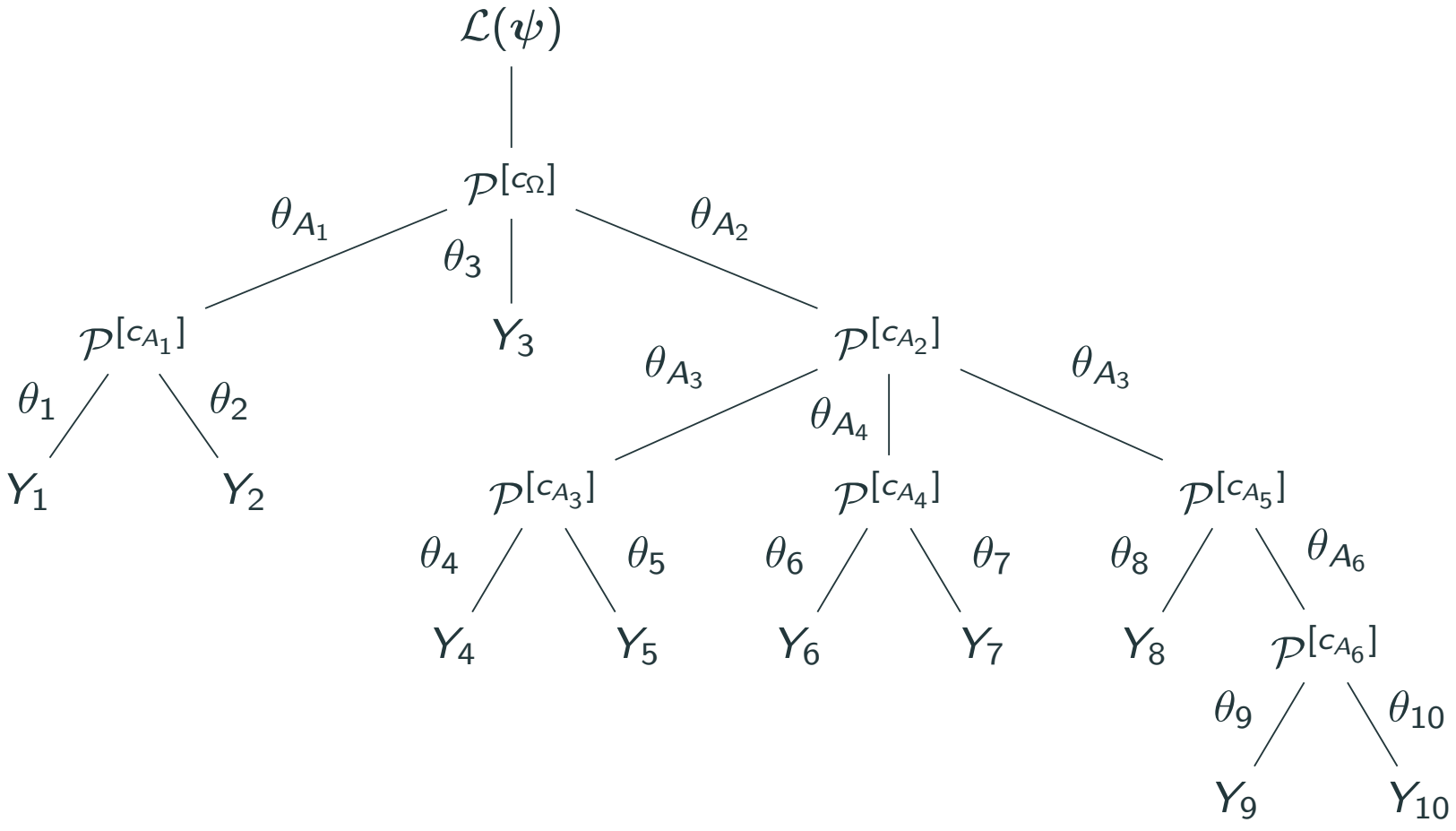


$$\mathfrak{C}_{A_2} = \{A_3, A_4, A_5\}$$

Tree Pólya Splitting - Definitions



Tree Pólya Splitting - Definitions



Tree Pólya Splitting - Definitions

Definition

The random vector $\mathbf{Y}$ has a *Tree Pólya Splitting* distribution with parameters $\mathfrak{T}$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_A)_{A\in\mathfrak{I}}$, $\mathbf{c} = (c_A)_{A\in\mathfrak{I}}$, and ψ if its p.m.f. is given by

$$\mathbb{P}(\mathbf{Y}=\mathbf{y}) = \underbrace{\mathbb{P}(|\mathbf{Y}|=|\mathbf{y}|)}_{\mathcal{L}(\psi)} \prod_{A\in\mathfrak{I}} \underbrace{\frac{n_A!}{(|\boldsymbol{\theta}_A|)_{(n_A,c_A)}} \prod_{C\in\mathfrak{C}_A} \frac{(\theta_C)_{(n_C,c_A)}}{n_C!}}_{\mathcal{P}^{[c_A]}(\boldsymbol{\theta}_A)},$$

where $\mathbf{Y}_A = (Y_j)_{j\in A}$ and $n_A := |\mathbf{y}_A|$ for any node $A \in \mathfrak{I}$.

Tree Pólya Splitting - Definitions

Definition

Such a distribution is denoted by

$$\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi}).$$

As an example, let us suppose $\mathcal{L} = \mathcal{NB}(\alpha, p)$ with $\alpha > 0$, $p \in (0, 1)$, and p.m.f.

$$\mathbb{P}(|\mathbf{Y}| = n) = \frac{(\alpha)_n}{n!} p^n (1-p)^\alpha.$$

Moreover, let us use only multinomial ($\mathcal{M}$) or Dirichlet-multinomial ($\mathcal{DM}$) distributions in $\mathfrak{T}$.

Tree Pólya Splitting - Definitions

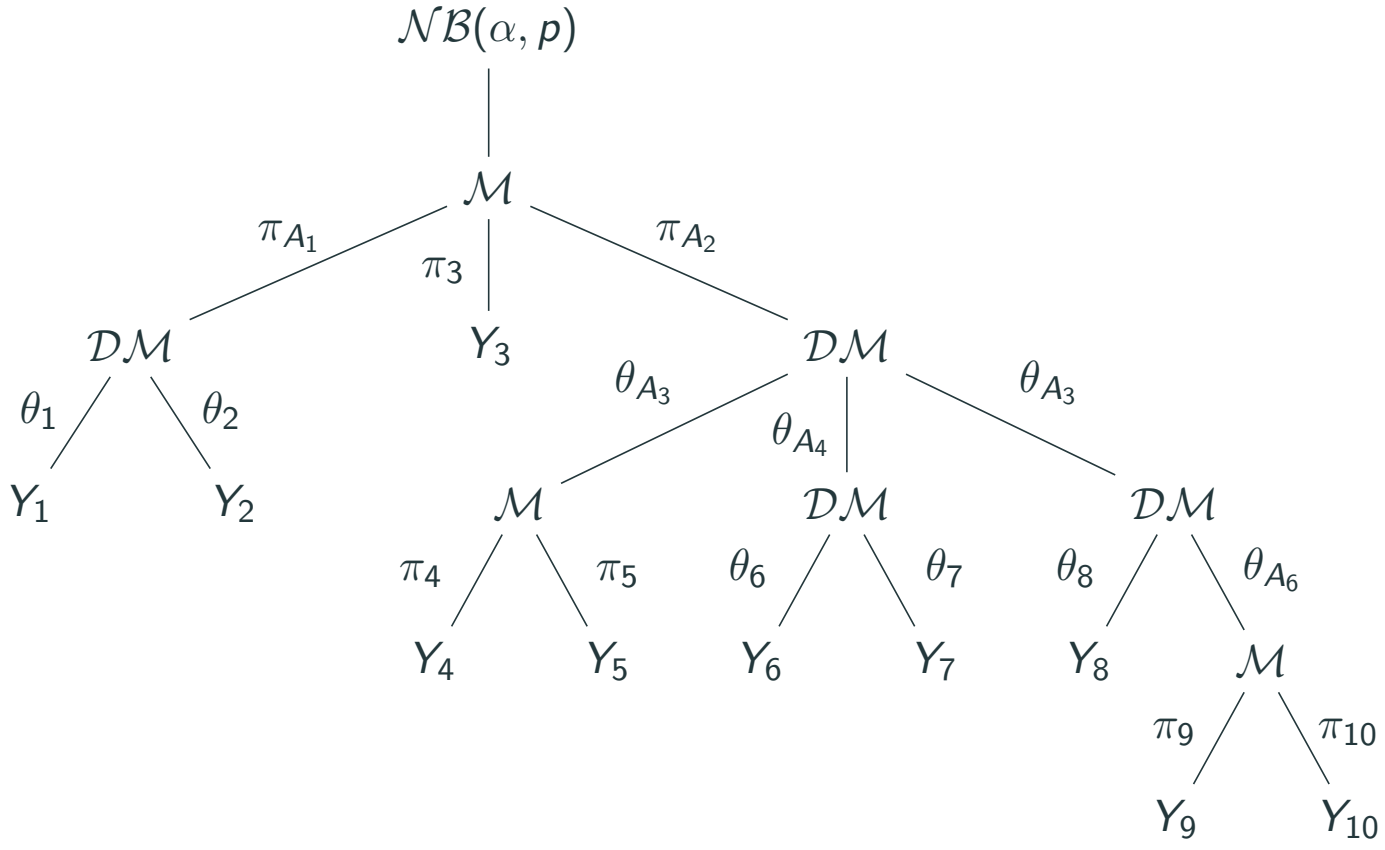


Figure 2: Exemple of a Tree Pólya Splitting model

Tree Pólya Splitting - Marginal

Theorem [Valiquette et al. (2025)]

Let $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$. The marginal distribution of Y_j is given by

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

where $A_k \in \mathrm{Path}_{\{j\}}$ and $\boldsymbol{\theta}_{A_k \setminus A_{k-1}}$ is the vector of parameters associated with the node A_k, such that $\theta_{A_{k-1}}$ is removed.

Tree Pólya Splitting - Marginal

Theorem [Valiquette et al. (2025)]

Let $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and $A \in \mathfrak{I}$. The marginal distribution of $|\mathbf{Y}_A|$ is given by (1), where $A_k \in \mathrm{Path}_A$.

If $A = \Omega$, then $K = 0$ and we retrieve $|\mathbf{Y}| \sim \mathcal{L}(\boldsymbol{\psi})$.

Tree Pólya Splitting - Marginal

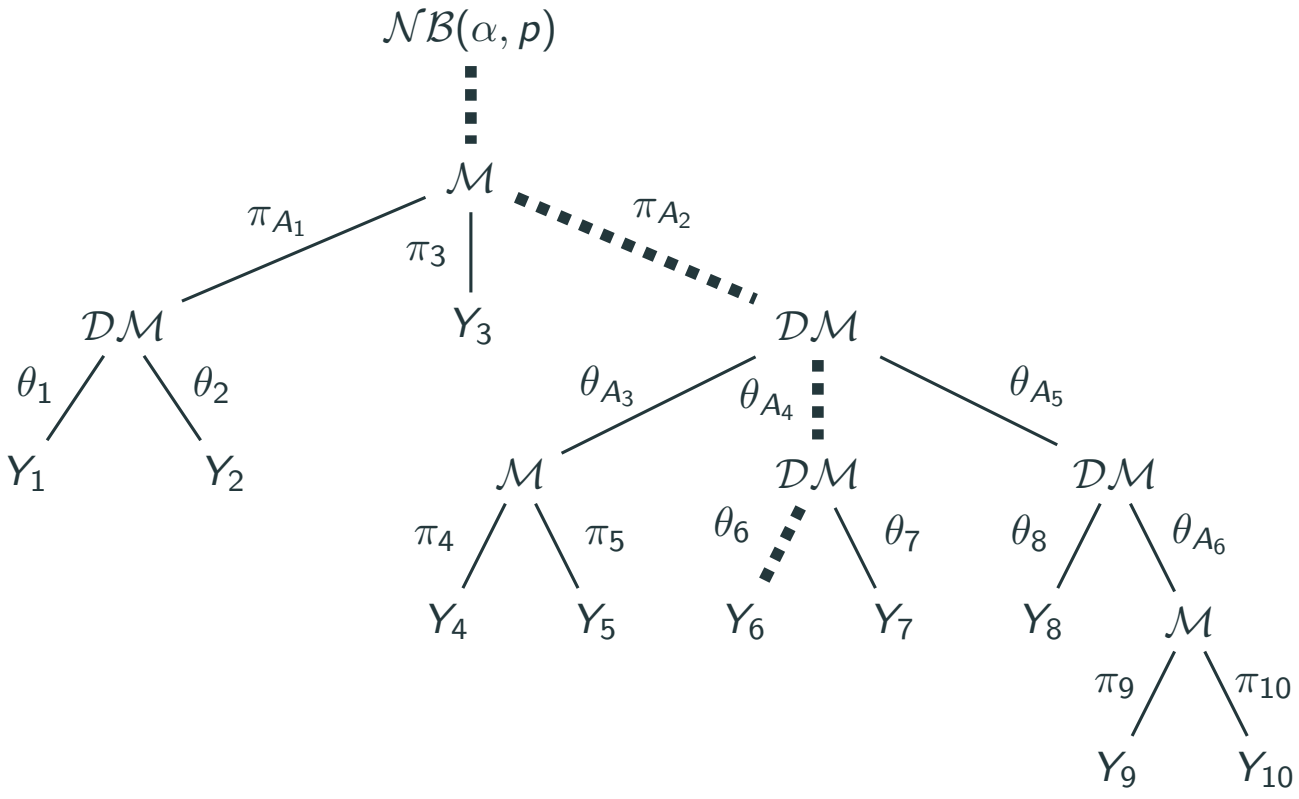
Theorem [Valiquette et al. (2025)]

Let $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\psi)$ and $A \in \mathfrak{I}$. The multivariate marginal distribution of $\mathbf{Y}_A$ is given by

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

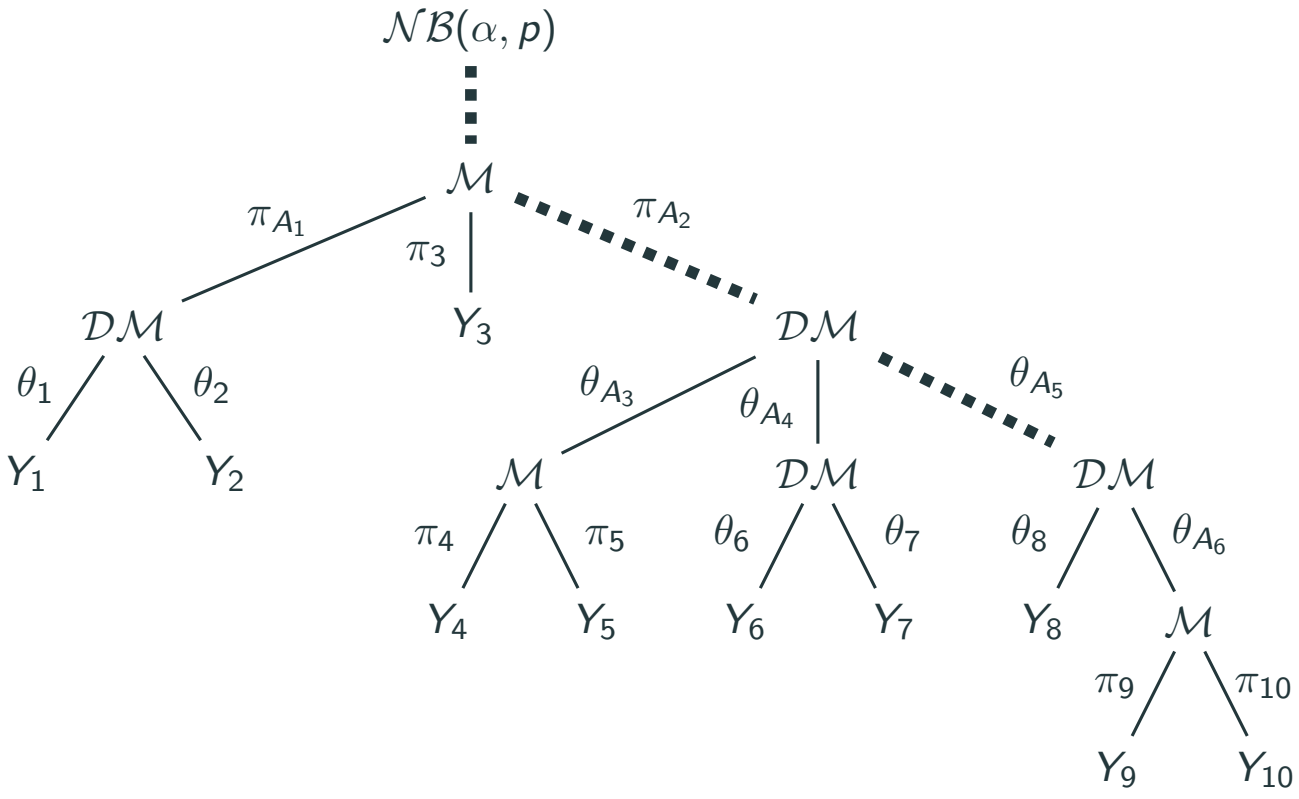
where $\tilde{\mathfrak{T}}$ is the pruned tree with root A, interior nodes $\tilde{\mathfrak{I}}$, leaves $\tilde{\mathfrak{L}}$, and parameters $\tilde{\boldsymbol{\theta}} = \{\boldsymbol{\theta}_A\}_{A \in \tilde{\mathfrak{I}}}$, $\tilde{\mathbf{c}} = \{\mathbf{c}_A\}_{A \in \tilde{\mathfrak{I}}}$.

Tree Pólya Splitting - Marginal



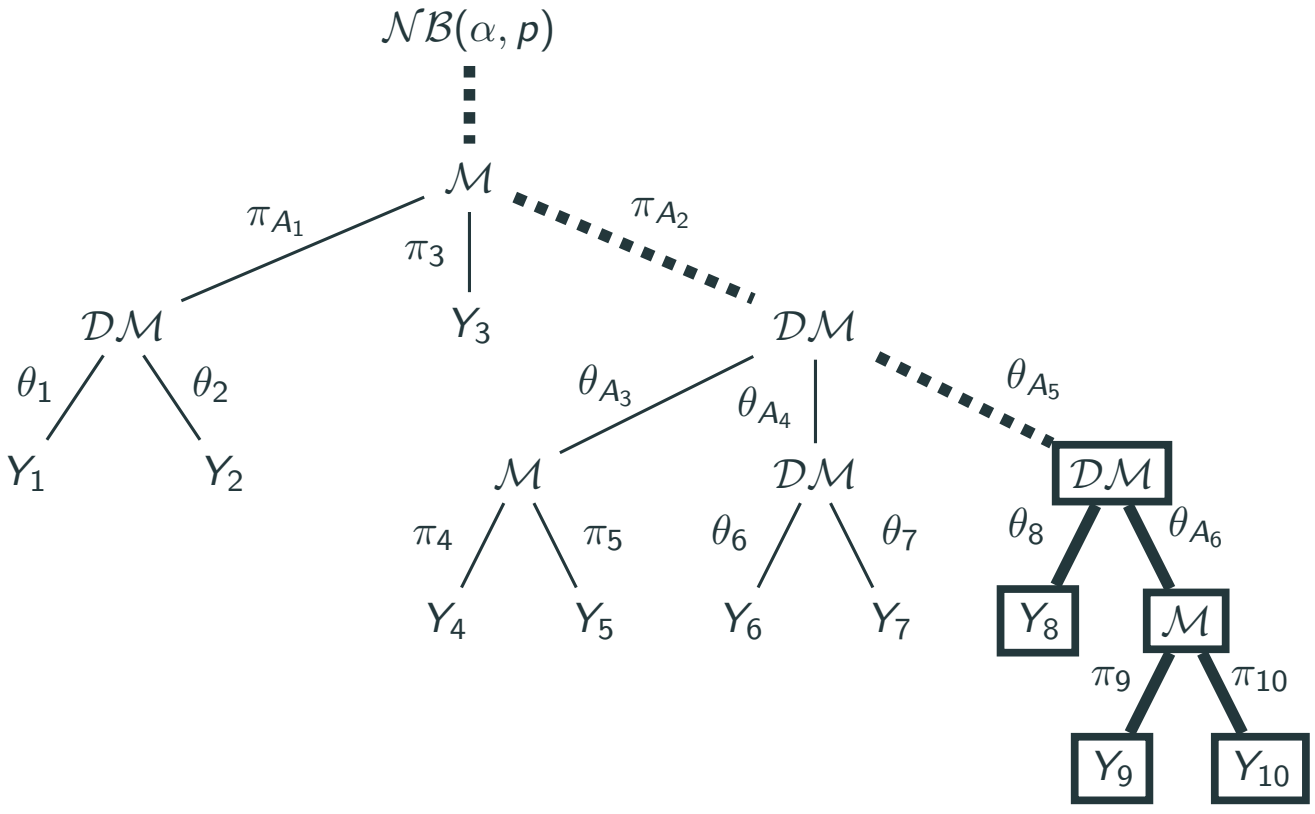
$$Y_6 \sim \mathcal{BB}_{n_1}(\theta_6, \theta_7) \underset{n_1}{\wedge} \mathcal{BB}_{n_2}(\theta_{A_4}, \theta_{A_3} + \theta_{A_5}) \underset{n_2}{\wedge} \mathcal{B}_{n_3}(\pi_{A_2}) \underset{n_3}{\wedge} \mathcal{NB}(\alpha, p)$$

Tree Pólya Splitting - Marginal



$$Y_8 + Y_9 + Y_{10} \sim \mathcal{BB}_{n_1}(\theta_{A_5}, \theta_{A_3} + \theta_{A_4}) \underset{n_1}{\wedge} \mathcal{B}_{n_2}(\pi_{A_2}) \underset{n_2}{\wedge} \mathcal{NB}(\alpha, p)$$

Tree Pólya Splitting - Marginal



$$(Y_8, Y_9, Y_{10}) \sim \mathcal{TP}_{\Delta_n}(\tilde{\mathfrak{T}}; \tilde{\boldsymbol{\theta}}, \tilde{\mathbf{c}}) \underset{n}{\wedge} (Y_8 + Y_9 + Y_{10})$$

Tree Pólya Splitting - Marginal

In this example, we can prove that all the distributions $\mathcal{B}_n$ can be "ignored". Therefore, when it comes to analyzing the marginal here, all of them have the form

$$X \sim \left[\bigwedge_{k=1}^{K} \mathcal{BB}_{n_k}(\alpha_k, \beta_k) \right] \underset{n_K}{\wedge} \mathcal{NB}(\alpha, p). \tag{2}$$

Tree Pólya Splitting - Marginal

For $K = 1$, Jones & Marchand [2019] prove that the p.m.f. of (2) is given by

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

where ${}_pF_q\left[\begin{matrix}\mathbf{a}\\ \mathbf{b}\end{matrix}; z\right]$ is the *generalized hypergeometric function* with $\mathbf{a} \in \mathbb{R}^p$, $\mathbf{b} \in \mathbb{R}_+^q$.

Tree Pólya Splitting - Marginal

Theorem [Valiquette et al. (2025)]

Let $p \in (0,1)$, $\alpha > 0$, $\boldsymbol{a} = (a_1, \ldots, a_K)$ and $\boldsymbol{b} = (b_1, \ldots, b_K)$ two positive vectors with $K \geq 1$. If X is distributed as in (2), then its p.m.f. is given by

$$\left[\prod_{k=1}^{K} \frac{(\alpha_k)_n}{(\alpha_k + \beta_k)_n}\right] \frac{(\alpha)_n}{n!} \left(\frac{p}{1-p}\right)^n \sum_{m=0}^{\infty} \frac{(\alpha+n)_m}{m!} p^m (1-p)^{\alpha+n} {}_{K+1}F_K \left[\begin{matrix} -m, \ \boldsymbol{a} + n\mathbf{1} \\ \boldsymbol{a} + \boldsymbol{b} + n\mathbf{1} \end{matrix}; 1\right],$$

where $\mathbf{1}$ is the unit vector.

Tree Pólya Splitting - Marginal

Theorem [Valiquette et al. (2025)]

In particular, if $p \in (0, 1/2)$, its p.m.f. is given by

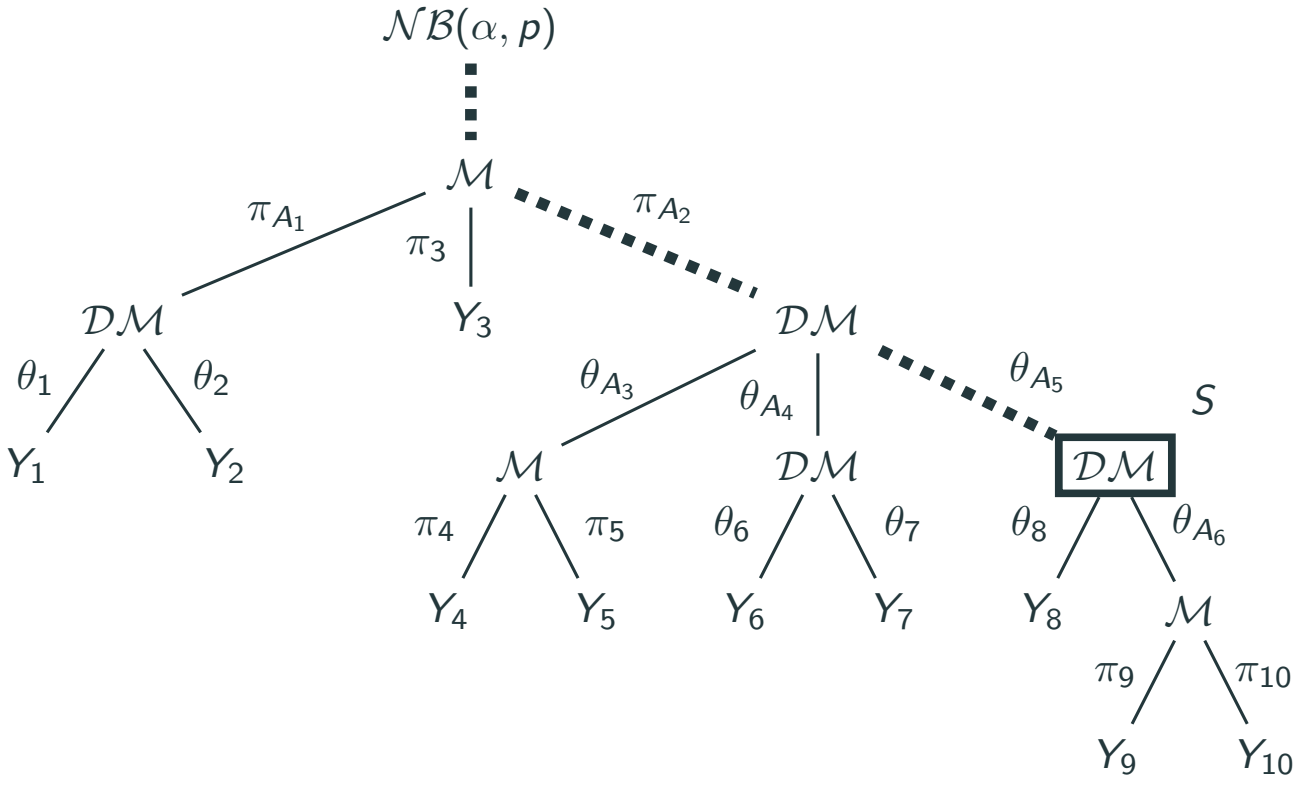
$$\left[\prod_{k=1}^{K} \frac{(\alpha_k)_n}{(\alpha_k + \beta_k)_n}\right] \frac{(\alpha)_n}{n!} \left(\frac{p}{1-p}\right)^n {}_{K+1}F_K\left[\begin{matrix} \alpha + n,\ \boldsymbol{a} + n\mathbf{1} \\ \boldsymbol{a} + \boldsymbol{b} + n\mathbf{1} \end{matrix}; \frac{p}{p-1}\right].$$

Tree Pólya Splitting - Covariance & correlation

Let $A \in \mathfrak{I} \cup \mathfrak{L}$ and $A_k \in \mathrm{Path}_A$, we define the following constants.

$$\gamma_A = \prod_{k=1}^{K} \frac{\theta_{A_{k-1}}}{|\boldsymbol{\theta}_{A_k}|} \text{ and } \delta_A = \prod_{k=1}^{K} \frac{\theta_{A_{k-1}} + c_{A_k}}{|\boldsymbol{\theta}_{A_k}| + c_{A_k}}.$$

Tree Pólya Splitting - Covariance & correlation



$$\gamma_S = \frac{\theta_{A_5}}{\theta_{A_3}+\theta_{A_4}+\theta_{A_5}} \cdot \pi_{A_2} \text{ and } \delta_S = \frac{\theta_{A_5}+1}{\theta_{A_3}+\theta_{A_4}+\theta_{A_5}+1} \cdot \pi_{A_2}$$

Tree Pólya Splitting - Covariance & correlation

Theorem [Valiquette et al. (2025)]

Let $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and marginals Y_i, Y_j $(i \neq j)$. Then, there is a common ancestor node $S \in \mathfrak{I}$ such that

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

Note: The sign only depends on S, which is dependent on the choice of Y_i and Y_j.

Tree Pólya Splitting - Covariance & correlation

Theorem [Valiquette et al. (2025)]

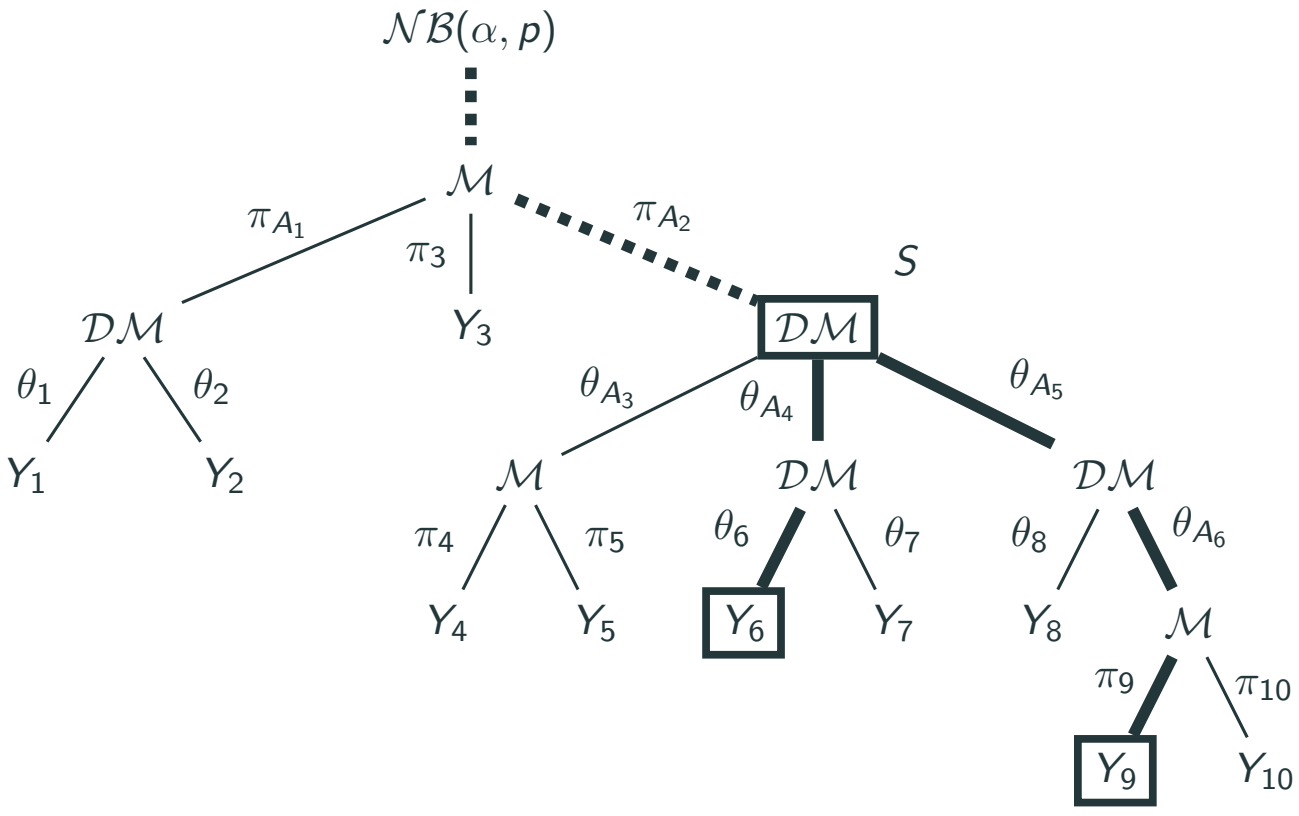
Let $\mathbf{Y} \sim \mathcal{TP}_{\Delta_n}(\mathfrak{T}; \boldsymbol{\theta}, \mathbf{c}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ and marginals Y_i, Y_j $(i \neq j)$. Then, there is a common ancestor node $S \in \mathfrak{I}$ such that

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

where

$$\Lambda_k = \sqrt{\frac{\gamma_\ell}{\delta_\ell \mu_2 + \mu_1 (1 - \gamma_\ell \mu_1)}};\ k = i,j.$$

Tree Pólya Splitting - Covariance & correlation



$$\mathrm{Cov}(Y_6, Y_9) = \alpha \left(\frac{p}{1-p}\right)^2 \left(\frac{|\boldsymbol{\theta}_S| - \alpha}{|\boldsymbol{\theta}_S| + 1}\right) \gamma_6 \gamma_9$$

Tree Pólya Splitting - Covariance & correlation

In this example, the sum at S is such that

$$|\mathbf{Y}_S| \sim \mathcal{NB}\left(\alpha, \frac{p\pi_{A_2}}{1-p+p\pi_{A_2}}\right),$$

and (Y_6, Y_9) are independent if and only if $\alpha = \theta_{A_3} + \theta_{A_4} + \theta_{A_5}$ (Peyhardi et al. [2021]).

Tree Pólya Splitting - Covariance & correlation

In fact, the sets $\{Y_4, Y_5\}$, $\{Y_6, Y_7\}$, and $\{Y_8, Y_9, Y_{10}\}$ are mutually independent if and only if

$$\alpha = \theta_{A_3} + \theta_{A_4} + \theta_{A_5},$$

since all their common ancestor is the node S.

Tree Pólya Splitting - Covariance & correlation

Let us fix the following parameters:

$$\boldsymbol{\alpha = 10}$$

$$p = 0.95$$

$$\pi_{A_1} = \pi_9 = 0.3$$

$$\pi_3 = 0.1$$

$$\theta_1 = \theta_2 = 1.5$$

$$\pi_{A_2} = 0.6$$

$$\boldsymbol{\theta_{A_3} = 3}$$

$$\boldsymbol{\theta_{A_4} = \theta_{A_5} = 3.5}$$

$$\pi_4 = \pi_5 = 0.5$$

$$\theta_6 = 0.8$$

$$\theta_7 = \theta_8 = 1$$

$$\theta_{A_6} = 2.5$$

$$\pi_{10} = 0.7$$

Tree Pólya Splitting - Covariance & correlation

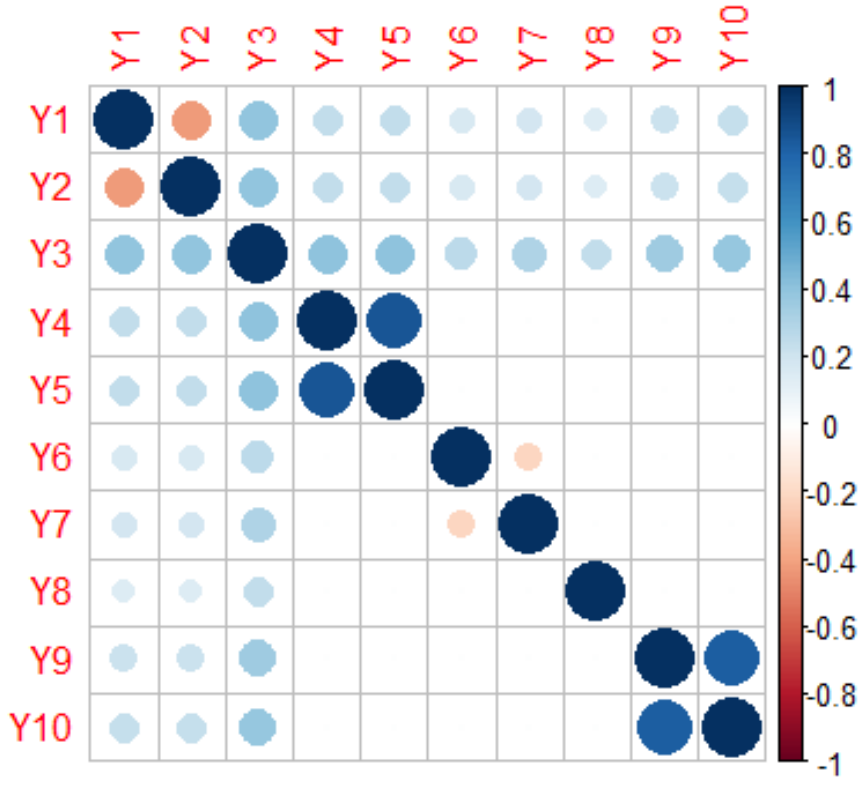


Figure 3: Correlation plot of the example

Application to Trichoptera dataset

Application to Trichoptera dataset

Tree Pólya Splitting is adjusted to the **Trichoptera dataset** provided by Usseglio-Polatera & Auda [1987]. It consists of $\mathbf{J = 17}$ **species'** abundances on $\mathbf{n = 49}$ **sites** between 1959 and 1960.

Application to Trichoptera dataset

Che	Hyc	Hym	Hys	Psy	Aga	Glo	Ath	Cea	Ced	Set
0	0	5	0	17	0	0	0	0	2	0
0	0	3	0	8	0	0	0	0	0	0
0	0	1	0	32	0	0	0	0	0	0
0	0	3	0	176	4	0	0	0	1	0
0	0	4	0	69	2	0	0	0	0	0
0	0	2	0	14	1	0	0	0	0	0

Table 1: Trichoptera dataset

Application to Trichoptera dataset

Using the AIC criterion, we built $\mathfrak{T}$ using the dataset and compared our model to other **Pólya Splitting** distributions and to the **multivariate Poisson-lognormal** (Chiquet et al. [2021]).

Application to Trichoptera dataset

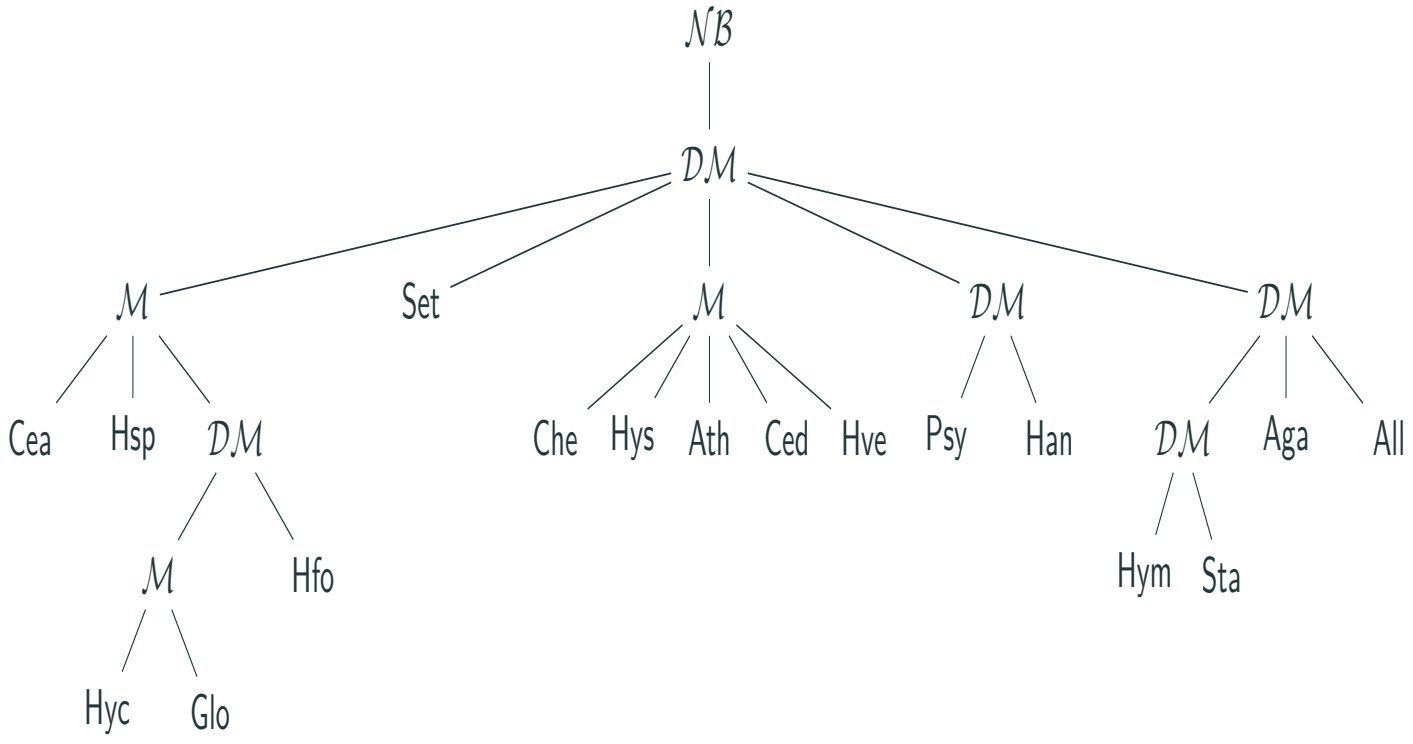


Figure 4: Tree Pólya Splitting fitted to the Trichoptera data.

Application to Trichoptera dataset

Models	Nb. Parameters	AIC
Tree Pólya Splitting	23	2380.77
Generalized Dirichlet-multinomial Splitting	34	2460.70
Dirichlet-multinomial Splitting	19	2494.87
Multivariate Poisson-lognormal	170	2599.63
Multinomial Splitting	18	6362.20

Table 2: Fitted models to the Trichoptera data.

Application to Trichoptera dataset

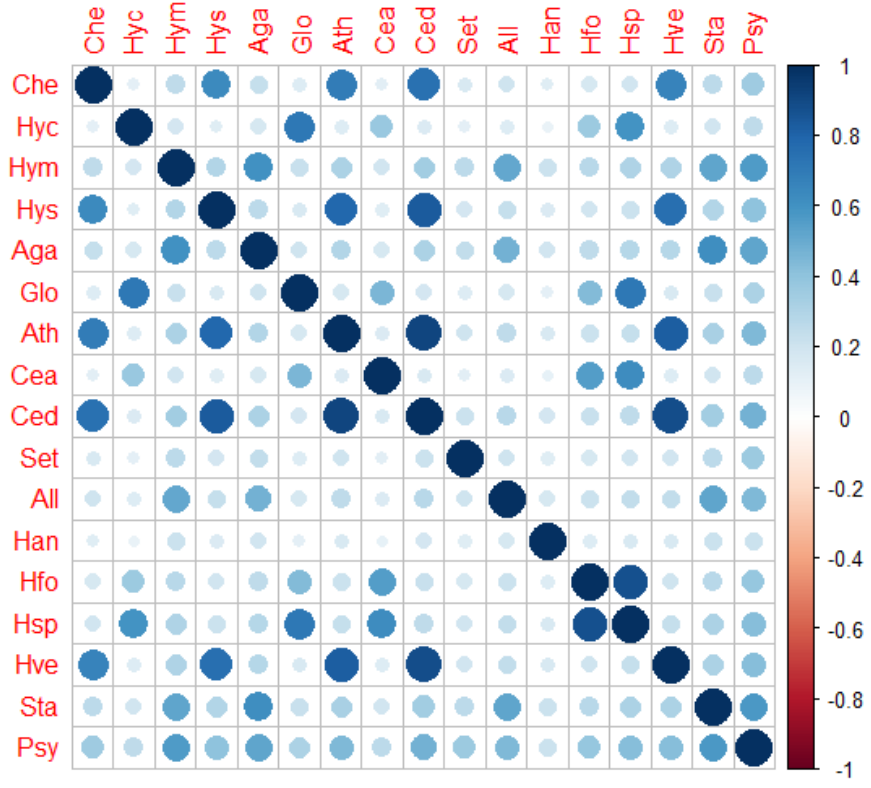


Figure 5: Correlation plot of the Tree Pólya Splitting fitted to the Trichoptera data.

Conclusion and perspectives

Conclusions and perspectives

Summary

- Tree Pólya Splitting is a flexible and simple model for multivariate count data.
- Its marginals, moments, correlations, and independences can be easily evaluated.

Perspectives

- Incorporating covariates at each internal node.
- Using the tree structure for multivariate zero inflation (Moudjeu et al. [2025]).
- Developing a probabilistic graphical model for such distributions.

Thank you for your attention!

Email: samuel.valiquette@mail.ca

Article: Tree Pólya Splitting distributions for multivariate count data (arXiv:2404.19528)

Bibliography

D. A. Aoudia, É. Marchand, and F. Perron (2016). **Counts of Bernoulli success strings in a multivariate framework**. *Statistics & Probability Letters*, 119, 1-10.

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

J. Chiquet, M. Mariadassou, and S. Robin (2021). **The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances**. *Frontiers in Ecology and Evolution*, 9:588292.

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

F. Moudjeu, J. Peyhardi, M. Réjou-Méchain, P. Soh Takam, and F. Mortier (2025). **Tree splitting regression models for multivariate count data: application to forest trees abundance** (To be submitted). *Biometrics*.

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

Bibliography

J. Peyhardi (2023). **On quasi Pólya thinning operator**. *Brazilian Journal of Probability and Statistics*, 37(4), 643-666.

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

P. Usseglio-Polatera, and Y. Auda (1987). **Influence des facteurs météorologiques sur les résultats de piégeage lumineux**. *Annales de Limnologie-International - Journal of Limnology*, 23(1), 65–79.

S. Valiquette, J. Peyhardi, É. Marchand, G. Toulemonde, and F. Mortier (2025). **Tree Pólya Splitting distributions for multivariate count data**. *Journal of Multivariate Analysis*.

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