Sur les données de comptage dans le cadre des valeurs extrêmes et la modélisation multivariée

Samuel Valiquette

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistique

École doctorale - Information, Structure, Systèmes (I2S)

Unité de recherche – IMAG

Équipe de recherche – GAMBAS

En partenariat international avec l'Université de Sherbrooke, CANADA

Sur les données de comptage dans le cadre des valeurs extrêmes et la modélisation multivariée

Présentée par Samuel VALIQUETTE

Le 3 juillet 2024

Sous la direction de Gwladys TOULEMONDE et Éric MARCHAND

Devant le jury composé de

Jean-Noël BACRO, Professeur, Université de Montpellier	Président-examinateur
Stéphane GIRARD, Directeur de recherche, Inria Grenoble Rhône-Alpes	Rapporteur
Klaus HERRMANN, Professeur adjoint, Université de Sherbrooke	Examinateur
Éric MARCHAND, Professeur, Université de Sherbrooke	Directeur de thèse
Frédéric MORTIER, Chargé de recherche, Cirad	Co-encadrant
Jean PEYHARDI, Maître de conférences, Université de Montpellier	Membre invité
Stéphane ROBIN, Professeur, Sorbonne Université	Rapporteur
Gwladys TOULEMONDE, Professeure, Université de Montpellier	Directrice de thèse

RÉSUMÉ

Cette thèse s'intéresse à certains aspects théoriques de la modélisation des données de comptage. Deux cadres distincts sont abordés : celui des valeurs extrêmes et celui de la modélisation multivariée. Notre première contribution explore, en termes des comportements extrêmes, les liens existants entre le mélange Poisson et sa loi de mélange. Ce travail permet de caractériser et séparer plusieurs familles de lois de mélanges Poisson selon leur comportement en queue. Bien que ce travail soit théorique, nous discutons aussi de son utilité d'un point de vue pratique, notamment pour le choix de la loi de mélange. Notre deuxième contribution porte sur une nouvelle classe de modèles multivariés dénommée *Tree Pólya Splitting*. Celle-ci repose sur une modélisation hiérarchique et suppose qu'une quantité aléatoire est répartie successivement selon une loi de Pólya à travers une structure d'arbre de partition. Dans ce travail, nous caractérisons les lois marginales univariées et multivariées, les moments factoriels, ainsi que les structures de dépendance (covariance/corrélation) qui en découlent. Nous mettons en évidence, à l'aide d'un jeu de données correspondant à l'abondance de trichoptères, l'intérêt de cette classe de modèles en comparant nos résultats à ceux obtenus, par exemple, avec des modèles de type Poisson log-normale multivariée. Nous concluons cette thèse en présentant diverses perspectives de recherche.

Mots clés : Données de comptage, modèle de distribution des espèces, valeurs extrêmes, modélisation multivariée discrète.

ABSTRACT

Title : On count data within the framework of extreme values and multivariate modeling.

This thesis focuses on theoretical aspects of counting data modeling. Two distinct frameworks are addressed : extreme values and multivariate modeling. Our first contribution explores, in terms of extreme behaviors, the existing connections between the Poisson mixture and its mixing distribution. This work allows us to characterize and discriminate several families of Poisson mixture according to their tail behavior. Although this work is theoretical, we discuss its practical utility, particularly regarding the choice of the mixing distribution. Our second contribution focuses on a new class of multivariate models called *Tree Pólya Splitting*. This class is based on hierarchical modeling and assumes that a random quantity is successively divided according to a Pólya distribution through a partition tree structure. In this work, we characterize univariate and multivariate marginal distributions, factorial moments, as well as the resulting dependence structures (covariance/correlation). Using a dataset corresponding to the abundance of Trichoptera, we highlight the interest of this class of models by comparing our results to those obtained, for example, with multivariate Poisson-lognormal models. We conclude this thesis by presenting various perspectives for future research.

Keywords : Count data, joint species distribution models, extreme values, discrete multivariate model.

REMERCIEMENTS

Dans un premier temps, j'aimerais remercier tout particulièrement mes superviseurs Éric Marchand, Frédéric Mortier, Jean Peyhardi et Gwladys Toulemonde. J'ai eu le privilège unique de collaborer avec chacun d'entre eux au cours des cinq dernières années. Leurs conseils et leurs expertises diverses m'ont permis d'élargir mes perspectives sur la statistique et la recherche. C'est grâce à leur encadrement et à leur générosité que ma confiance s'est développée, ainsi que ma détermination à poursuivre une carrière en recherche. C'est un grand honneur de présenter dans cette thèse le fruit de cette collaboration que je chérirai pour toujours. Ce travail n'aurait pas vu le jour sans eux. Je leur en suis donc très reconnaissant.

J'aimerais remercier sincèrement le *Centre de coopération internationale en recherche agronomique pour le développement* (CIRAD) et le projet *Generating Advances in Modeling Biodiversity And ecosystem Services* (GAMBAS) pour le financement de cette thèse. Cette aide m'a grandement soulagé, me permettant ainsi de me concentrer principalement sur mes travaux. Ils ont pour cela toute ma gratitude. J'aimerais également remercier chaque membre de ces équipes pour m'avoir accompagné tout au long de cette aventure. Un remerciement particulier à Fabrice Moudjieu pour nos nombreuses discussions ; ce fut un grand plaisir de travailler ensemble.

Je remercie l'*Institut national de recherche en sciences et technologies du numérique* (Inria) et l'équipe *Littoral, Environnement, Modèles et Outils Numériques* (LEMON)

pour l'accueil chaleureux et amical. Ce fut un plaisir de visiter le laboratoire à chacun de mes passages à Montpellier. J'aimerais particulièrement remercier Antoine Rousseau de m'avoir fait sentir inclus dans l'équipe.

Évidemment, cette cotutelle n'aurait pas lieu sans l'aide des écoles doctorales de l'Université de Sherbrooke et de l'Université de Montpellier. Je remercie chaque personne impliquée dans ce processus laborieux qui a rendu cette collaboration académique possible. Des remerciements particuliers à Joseph Salmon et Josée Lamoureux pour leur aide au bon déroulement de cette thèse. J'aimerais également remercier tous mes collègues doctorantes et doctorants que j'ai eu la chance de rencontrer au cours de cette thèse.

Je suis reconnaissant envers les rapporteurs Stéphane Girard et Stéphane Robin, ainsi que les examinateurs Jean-Noël Bacro et Klaus Herrmann, pour leurs commentaires constructifs et pour avoir accepté de participer à ma soutenance.

Je tiens à remercier le *Conseil de recherche en sciences naturelles et en génie du Canada* (CRSNG), l'*Institut des sciences mathématiques* (ISM) et l'Université de Sherbrooke pour leur soutien financier durant la quatrième année de cette thèse. En particulier, je suis reconnaissant envers l'ISM pour m'avoir accordé la bourse d'excellence ainsi que la subvention de voyage, ainsi qu'envers l'Université de Sherbrooke pour la bourse de valorisation.

Finalement, j'aimerais exprimer ma plus grande gratitude à mes parents, ma sœur et mes proches pour m'avoir accompagné tout au long de cette aventure. Vos encouragements et votre soutien depuis le début de mon parcours universitaire ont contribué à faire de moi la personne que je suis aujourd'hui. Cette thèse est dédiée à vous.

Samuel Valiquette

Sherbrooke, 3 juillet 2024

TABLE DES MATIÈRES

NOTATIONS ET ABRÉVIATIONS

$\mathbb{R}$, $\mathbb{R}_+$	Ensemble des nombres réels et des nombres réels strictement positifs, respectivement
$\mathbb{N}$, $\mathbb{N}_+$	Ensemble des nombres naturels et des nombres naturels excluant 0, respectivement
i.i.d.	Indépendantes et identiquement distribuées
$\mathrm{E}(X)$	Espérance de la variable aléatoire X
$\mathrm{Var}(X)$	Variance de la variable aléatoire X
$\mathrm{Cov}(X, Y)$	Covariance entre les variables aléatoires X et Y
$\mathrm{Corr}(X, Y)$	Corrélation de Pearson entre les variables aléatoires X et Y
$f(x) \sim g(x)$	$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$: équivalence asymptotique lorsque $x \to \infty$
$f(x) = o(g(x))$	$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$: relation petit o lorsque $x \to \infty$
$f(x) = O(g(x))$	$\limsup_{x\to\infty} \frac{f(x)}{g(x)} < \infty$: relation grand O lorsque $x \to \infty$
$\mathcal{RV}_\alpha$, $\mathcal{RV}_0$	Ensemble des fonctions à variation régulière et lente, res-

	pectivement
$\mathcal{D}_-$, $\mathcal{D}_0$, $\mathcal{D}_+$	Domaines d'attraction de Weibull, Gumbel et Fréchet, respectivement
$\mathcal{D}_0^{\mathcal{E}}$, $\mathcal{D}_0^{\mathcal{F}}$, $\mathcal{D}_0^{\mathcal{H}}$	Domaines d'attraction de Gumbel à queue exponentielle, finie et à défaillance Gumbel, respectivement
$\mathcal{P}$	Distribution de Poisson
$\mathcal{NB}$	Distribution binomiale négative
F_M, $\overline{F}_M$, P_M	Fonctions de répartition, de survie et de masse d'un mélange Poisson, respectivement
$\lvert\mathbf{x}\rvert$	$\sum_j x_j$: Norme L_1
$\mathbf{x}_{\mathcal{J}}$, $\mathbf{x}_{-\mathcal{J}}$	Sous-vecteurs indexés par $\mathcal{J}$ et son complément, respectivement
Δ	$\{\mathbf{x} \in \mathbb{R}_+^J : \lvert\mathbf{x}\rvert = 1\}$: simplexe continu
Δ_n	$\{\mathbf{x} \in \mathbb{N}_+^J : \lvert\mathbf{x}\rvert = n\}$: simplexe discret
$(x)_{(n,c)}$	$\prod_{j=0}^{n-1}(x + cj)$: factorielle généralisée
$(x)_n$	$(x)_{(n,1)}$: factorielle ascendante
$(\mathbf{x})_n$	$\prod_{j=1}^{J}(x_j)_n$
$(\mathbf{x})_{\mathbf{r}}$	$\prod_{j=1}^{J}(x_j)_{r_j}$
$\wedge$, $\bigwedge_{j=1}^{J}$	Opérateurs de mélange et mélange itéré, respectivement
$\mathcal{P}_{\Delta_n}^{[c]}$	Distribution Pólya

$\mathcal{TP}_{\Delta_n}$	Distribution Tree Pólya
$\mathcal{M}_{\Delta_n}$	Distribution multinomiale
$\mathcal{DM}_{\Delta_n}$	Distribution Dirichlet-multinomiale
$\mathcal{H}_{\Delta_n}$	Distribution hypergéométrique
$\mathcal{L}$	Distribution discrète univariée dans les modèles Pólya Splitting et Tree Pólya Splitting
$\mathcal{D}_n$	Distribution de Dirac
$\mathcal{B}_n$	Distribution binomiale
$\mathcal{BB}_n$	Distribution bêta-binomiale
$\mathfrak{T}$	Arbre de partition
Ω	Nœud racine
$\mathfrak{I}$	Ensemble de nœuds internes
$\mathfrak{L}$	Ensemble de feuilles
$\mathfrak{C}_A$	Ensemble d'enfants du nœud A
$\mathscr{P}(A)$	Nœud parent de A
Path_A^B, Path_A	Chemins du nœud A aux nœuds B et Ω, respectivement

INTRODUCTION

En 1666, l'intendant Jean Talon produit le premier recensement de l'histoire du Canada afin d'améliorer le développement de la Nouvelle-France. Pour ce faire, M. Talon dénombre les habitants. Quatre siècles plus tard, comprendre l'impact des changements climatiques sur l'abondance des espèces est devenu un enjeu majeur [cf. Wenger et Freeman, 2008; Ovaskainen et Soininen, 2011; Ovaskainen et Abrego, 2020]. Les données de comptage étaient et sont toujours essentielles au progrès et à notre compréhension. Comme en écologie, différents domaines de recherche s'intéressent également à analyser ces données et à inférer certaines caractéristiques du phénomène qui les a générées. Le nombre de réclamations en actuariat [cf. Bartoszewicz, 2005; Boucher *et al.*, 2008] ou le nombre de mutations au sein des génomes en microbiologie [cf. Anders et Huber, 2010; Chen et Li, 2013; Kaul *et al.*, 2017] sont autant d'exemples qui reflètent l'importance des données de comptage. Toutefois, la qualité des analyses reste dépendante des modèles utilisés et donc de leurs propriétés mathématiques. D'un point de vue formel, les données de comptage peuvent être appréhendées comme la réponse de variables aléatoires discrètes. D'un point de vue probabiliste, celles-ci semblent simples et intuitives dans le sens où les mesures associées et les espaces de probabilités qui en découlent sont relativement intuitifs au regard des variables continues. Paradoxalement, le cadre continu apparaît, d'un point vue pratique, plus simple. Deux raisons principales peuvent être mentionnées : le cadre continu (i) a été de manière évidente nettement plus étudié et *de*

facto plus développé, (ii) repose sur des propriétés qui peuvent ne pas être vérifiées dans le cas discret. Prenons ici deux exemples parmi d'autres mais que nous développons dans cette thèse : la modélisation des valeurs extrêmes discrètes et la modélisation des lois multivariées discrètes.

La théorie des valeurs extrêmes permet de caractériser le comportement en queue d'une distribution à l'aide de trois domaines d'attraction : Weibull, Gumbel et Fréchet [cf. Resnick, 1987; Coles, 2001]. Alors que la plupart des distributions continues courantes peuvent être associées à un domaine d'attraction, ce n'est pas toujours le cas pour les variables aléatoires discrètes. Anderson [1970] démontre qu'il est nécessaire, pour être dans un domaine d'attraction, qu'une loi discrète ait une queue de distribution suffisamment lourde. Cette propriété n'est pas vérifiée pour les lois usuelles comme la loi de Poisson, la loi géométrique ou encore la loi binomiale négative. Néanmoins, Anderson [1970] décrit les comportements en queue de ces distributions et propose des approximations raisonnables du comportement en loi des valeurs extrêmes associées.

La modélisation de lois multivariées peut s'appréhender, grâce aux travaux de Sklar [1959], par l'étude des lois marginales et des copules. Plus précisément, pour une distribution conjointe continue de dimension J, Sklar démontre qu'il existe une copule unique décrivant les relations de dépendance entre les marginales. Dans ce contexte, ce résultat est un outil puissant de modélisation. Toutefois, l'unicité de la copule n'est plus vérifiée lorsque les marginales sont discrètes rendant ainsi ces modèles non identifiables [cf. Genest et Nešlehová, 2007]. Il est donc préférable de travailler directement avec les familles de distributions jointes discrètes. Puisque la distribution univariée de Poisson apparaît naturellement dans les processus de comptage, il semble raisonnable d'utiliser des distributions multivariées dérivées de celle-ci [cf. Inouye *et al.*, 2017]. Campbell [1934] et Teicher [1954] proposent l'utilisation de $J+1$ variables aléatoires de Poisson indépendantes afin de créer une distribution multivariée de J Poisson dépendantes. Malgré la simplicité

de leur approche, la distribution résultante possède nécessairement des corrélations positives. Cependant, les relations de dépendance sont généralement plus diversifiées que simplement des dépendances positives. De plus, les variables discrètes présentent souvent une surdispersion, c'est-à-dire que leurs variances excèdent leurs espérances. Une solution proposée par Aitchison et Ho [1989] et appliquée en écologie par Chiquet *et al.* [2021] est la distribution Poisson log-normale multivariée. Cette dernière possède non seulement des marginales surdispersées, mais également une structure de corrélation flexible. Toutefois, il est important de noter que ces dépendances sont au niveau de l'espace latent et non au niveau des observations. Plus précisément, une corrélation nulle n'implique pas systématiquement une indépendance entre les observations. Il serait donc intéressant d'obtenir un modèle multivarié discret qui capture la véritable structure de dépendance tout en préservant une certaine flexibilité comme celle obtenue dans le cadre de la distribution Poisson log-normale multivariée.

Cette thèse vise à apporter de nouvelles perspectives et contributions à ces deux questions théoriques. La première partie porte sur les modèles de mélange Poisson et leurs comportements extrêmes. Ce cadre méthodologique est efficace pour modéliser des données de comptage présentant de la surdispersion qui peut être induite soit par des évènements extrêmes, soit par un excès de zéros [cf. Karlis et Xekalaki, 2005]. Par définition, une variable aléatoire discrète est distribuée selon un mélange Poisson si le paramètre de la loi de Poisson est elle-même une variable aléatoire réelle positive. Or ces mélanges sont des distributions discrètes, elles sont donc soumises aux conditions établies par Anderson [1970]. Néanmoins, la loi de mélange peut facilement bénéficier de la théorie des valeurs extrêmes classique lorsqu'elle est continue. Puisque cette dernière caractérise le mélange Poisson [cf. Feller, 1943], il serait intéressant de comprendre la relation entre son domaine d'attraction et le comportement en queue du mélange Poisson. Perline [1998] explore cette question et présente deux situations où la loi de mélange permet à la distribution dis-

crète de posséder un domaine d'attraction. Toutefois, Perline reconnaît que ses résultats n'établissent aucune connexion avec les travaux d'Anderson et laisse cette question non résolue. Willmot [1990] examine également cette interaction, mais sous un angle différent. En particulier, il s'intéresse aux relations entre les comportements asymptotiques de la densité de mélange et de la fonction de masse du mélange résultant. Il démontre, entre autres, que si cette densité se comporte asymptotiquement comme une distribution gamma, alors la fonction de masse se comporte comme une binomiale négative à l'infini. À la lumière de ces résultats, il serait intéressant d'étudier cette relation pour d'autres types de densités. Est-il envisageable d'obtenir des équivalences similaires pour d'autres comportements asymptotiques ?

La deuxième partie de cette thèse s'intéresse à la modélisation probabiliste des lois multivariées de comptage en portant son attention sur l'extension des lois dîtes *Splitting*. Celles-ci, initialement proposées indépendemment par Peyhardi et Fernique [2017] et Jones et Marchand [2019], et généralisées par Peyhardi *et al.* [2021], reposent sur la modélisation jointe de la somme des composantes d'un vecteur d'abondance et de sa répartition au sein de chacune de ses composantes. En particulier, cette distribution se nomme la *Pólya Splitting* lorsque cette division aléatoire est déterminée par la loi de Pólya [Eggenberger et Pólya, 1923]. Cette classe de distributions contient de nombreuses lois connues, comme la multinomiale, la négative multinomiale, la Dirichlet-multinomiale, mais aussi des lois moins usitées en écologie comme la loi de Waring généralisée [Xekalaki, 1986] ou la loi de Schur-constante discrète [Castañer *et al.*, 2015]. Cette famille de lois s'avère particulièrement intéressante à plus d'un titre. Par exemple, il est possible de caractériser les dépendances au niveau des espèces [Peyhardi, 2023], de caractériser les graphes de dépendance [Peyhardi *et al.*, 2021] ou encore de mieux comprendre les lois stationnaires des processus de naissance et mort à temps discret utilisées dans le cadre de la théorie neutre en écologie [Peyhardi *et al.*, 2024]. Toutefois, ces modèles restent

rigides, notamment en terme de structure de dépendance. Est-il possible de généraliser ces modèles afin de remédier à cette contrainte ?

Les travaux présentés dans cette thèse tentent de répondre à ces diverses questions. Plus spécifiquement, au Chapitre 2, nous présentons notre article publié dans le journal *Stat* sur les comportements en queue des mélanges Poisson selon la loi de mélange [cf. Valiquette *et al.*, 2023]. Ce travail présente plusieurs familles distinctes de mélanges Poisson, chacune ayant l'une des propriétés suivantes : 1) être dans le domaine d'attraction Gumbel ou Fréchet ; 2) avoir un comportement extrême pouvant être approché par la Gumbel ; 3) ne permettre aucune approximation de ses extrêmes. Dans le premier cas, nous généralisons le résultat de Perline [1998] concernant le domaine Fréchet. En effet, celui-ci suppose plusieurs hypothèses sur la densité de mélange afin que le mélange Poisson soit dans le domaine Fréchet. Nous démontrons que ce résultat est vérifié si loi de mélange est simplement dans le domaine de Fréchet. Les deux autres situations permettent d'établir la connexion avec les résultats d'Anderson [1970]. Plus précisément, nous démontrons que si la loi de mélange possède une décroissance exponentielle ou une queue finie, alors les situations 2) et 3) sont respectivement atteintes. Autrement, nous obtenons un résultat complémentaire à ceux de Willmot [1990]. En effet, nous démontrons que si la loi de mélange est dans le domaine d'attraction de Weibull, alors la fonction de masse du mélange Poisson se comporte asymptotiquement comme une distribution Poisson. Nous complétons ce travail en évaluant la qualité de l'ajustement du mélange Poisson selon son comportement en queue, à la fois de manière analytique et numérique. Cette analyse s'avère particulièrement utile pour choisir la loi de mélange.

Dans le Chapitre 3, nous proposons une généralisation du modèle Pólya Splitting que nous nommons *Tree Pólya Splitting*. Ce modèle utilise un principe similaire où la somme des composantes du vecteur d'abondance est divisé successivement selon des distributions Pólya Splitting le long d'un arbre de partition. Sur la base des travaux de Peyhardi *et al.*

[2021], nous définissons rigoureusement cette distribution et présentons diverses propriétés similaires à celles de la distribution Pólya Splitting. Nous étudions, entre autres, les lois marginales univariées et multivariées, les moments factoriels, ainsi que les structures de dépendance qui en découlent. En particulier, nous démontrons que ce nouveau modèle possède une corrélation beaucoup plus flexible permettant de mieux capturer la véritable structure de dépendance des données d'abondance. De plus, nous présentons quelques généralisations particulières aux travaux de Jones et Marchand [2019]. Plus précisément, nous obtenons une nouvelle borne pour la corrélation de leur modèle et généralisons la fonction de masse marginale qu'ils obtiennent. Enfin, nous complétons ce chapitre avec une application de ce modèle à des données d'abondance de trichoptères et le comparons, entre autres, avec la Poisson log-normale multivariée et le modèle Pólya Splitting. Nous proposons également une méthode pour construire l'arbre à partir des observations.

Nous concluons avec le Chapitre 4 en présentant plusieurs perspectives de recherches que nos contributions apportent. En particulier, nous explorons les représentations graphiques probabilistes des modèles Tree Pólya Splitting, l'incorporation des excès de zéros dans ce modèle, une extension de nos travaux sur les mélanges Poisson et une connexion intéressante entre les modèles Pólya Splitting et nos travaux au Chapitre 2.

CHAPITRE 1

Résultats fondamentaux

1.1 Fonctions à variation régulière

Afin de bien comprendre la théorie des valeurs extrêmes ainsi que les résultats présentés au Chapitre 2, on introduit les différentes relations asymptotiques et les fonctions à variation régulière ainsi que leurs propriétés. Les références principalement utilisées sont Bingham *et al.* [1987] et Resnick [1987].

Puisque la théorie des valeurs extrêmes étudie les comportements asymptotiques des distributions de probabilité, il est nécessaire de déterminer les différentes relations asymptotiques utilisées dans cette thèse. Soit deux fonctions positives f et g, alors f est asymptotiquement équivalente à g lorsque $x \to \infty$ si et seulement si $\lim_{x\to\infty} f(x)/g(x) = 1$. On note cette relation par $f(x) \sim g(x)$. Il est important pour le reste de cette thèse de ne pas confondre ce symbole entre deux fonctions et $\sim$ entre une variable aléatoire et sa distribution. On dit $f(x)$ est grand O de $g(x)$ lorsque $x \to \infty$, noté par $f(x) = O(g(x))$, si et seulement si $\limsup_{x\to\infty} f(x)/g(x) < \infty$. Finalement, on dit $f(x)$ est petit o de $g(x)$ lorsque $x \to \infty$, noté par $f(x) = o(g(x))$, si et seulement si $\lim_{x\to\infty} f(x)/g(x) = 0$. Ces

définitions sont similaires lorsque x tend vers une valeur finie quelconque.

Les fonctions dites à *variation régulière* sont essentielles pour obtenir les résultats fondamentaux de la théorie des valeurs extrêmes. Intuitivement, une fonction est à variation régulière si elle se comporte à l'infini comme une fonction puissance. Précisément, on a la définition suivante.

Définition 1.1 (Bingham *et al.* [1987], p. 18)**.** Une fonction mesurable $U : \mathbb{R}_+ \to \mathbb{R}_+$ est à variation régulière d'indice $\alpha \in \mathbb{R}$ si pour tout $t > 0$,

$$U(x) \sim t^{-\alpha} U(tx), \text{ lorsque } x \to \infty. \tag{1.1}$$

On note cette propriété $U \in \mathcal{RV}_\alpha$.

En particulier, lorsque $\alpha = 0$, la fonction est dite à *variation lente*. Trivialement, il existe une fonction $C \in \mathcal{RV}_0$ telle que pour $U \in \mathcal{RV}_\alpha$, $U(x) = C(x)x^\alpha$. Une fonction à variation régulière, et de même à variation lente, peut être formulée à l'aide de la *représentation de Karamata*.

Théorème 1.1 (Bingham *et al.* [1987], p. 21)**.** $U \in \mathcal{RV}_\alpha$ si et seulement si U est telle que

$$U(x) = c(x) \exp\left[\int_1^x t^{-1}\eta(t)dt\right] \tag{1.2}$$

pour $x > 1$, $c : \mathbb{R}_+ \to \mathbb{R}_+$ et $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ avec $\lim_{x\to\infty} c(x) = c < \infty$ et $\lim_{x\to\infty} \eta(x) = \alpha$.

Cette représentation est utile, entre autres, pour démontrer que la convergence à la Définition 1.1 est uniforme.

Théorème 1.2 (Bingham *et al.* [1987], p. 22)**.** Soit $U \in \mathcal{RV}_\alpha$ et $0 < a < b < \infty$, alors $U(x) \sim t^{-\alpha}U(tx)$ uniformément lorsque $x \to \infty$ pour tout

— $t \in [a, b]$;

— $t \in (0, b]$ pour $\alpha > 0$ et U bornée ;

— $t \in [a, \infty)$ pour $\alpha < 0$.

Cette représentation permet également de mieux comprendre le comportement asymptotique de $U \in \mathcal{RV}_\alpha$ selon l'indice α. En effet, lorsque $\alpha > 0$, la fonction U se comporte à l'infini comme une fonction monotone croissante. De manière similaire, U se comporte éventuellement comme une fonction monotone décroissante si $\alpha < 0$.

Théorème 1.3 (Bingham *et al.* [1987], p. 22-23)**.** a) Soit $U \in \mathcal{RV}_\alpha$, alors

$$\lim_{x\to\infty} U(x) = \begin{cases} 0 & \text{si } \alpha < 0, \\ \infty & \text{si } \alpha > 0. \end{cases}$$

b) Soit $C \in \mathcal{RV}_0$, une fonction localement bornée sur $[0, \infty)$ et $\alpha > 0$, alors lorsque $x \to \infty$:

— $\sup_{t\le x}\{t^\alpha C(t)\} \sim x^\alpha C(x)$;

— $\sup_{t\ge x}\{t^{-\alpha} C(t)\} \sim x^{-\alpha} C(x)$.

Deux résultats restent nécessaires pour la suite. Le premier correspond au théorème de densité monotone dans le cas des fonctions de survie $\overline{F}$ avec densité f. Il sera utile au Chapitre 2 afin de distinguer les différentes familles de mélanges Poisson. Le second correspond à un théorème de type abélien pour les fonctions à variation régulière [Vuilleumier, 1967]. Dans le contexte de cette thèse, nous proposons une version probabiliste simplifiée de ce dernier qui s'avère suffisante pour analyser les fonctions de survie des mélanges Poisson. Nous présentons la preuve de Vuilleumier [1967] adaptée pour nos travaux. Une généralisation est décrite dans Bingham *et al.* [1987] aux Théorèmes 4.1.4 et 4.2.1.

Théorème 1.4 (Bingham *et al.* [1987], p. 39)**.** Soit $\overline{F}(x) = \int_x^\infty f(t)dt$ avec densité de probabilité f décroissante pour tout $x > x_0$ pour un certain $x_0 < \infty$. De plus, supposons $\overline{F}(x) \sim C(x)x^{-\alpha}$ lorsque $x \to \infty$, où $\alpha > 0$ et $C \in \mathcal{RV}_0$. Alors $f(x) \sim C(x)\alpha x^{-\alpha-1}$ lorsque $x \to \infty$.

Théorème 1.5 (Vuilleumier [1967])**.** Soit Y_n une suite de variables aléatoires positives de densités f_n pour $n \in \mathbb{N}$. Supposons qu'il existe des constantes $\delta > 0$ et $M > 0$ indépendantes de n telles que $\mathrm{E}[Y_n^{\pm\delta}] \sim Mn^{\pm\delta}$ lorsque $n \to \infty$. Alors pour tout $C \in \mathcal{RV}_0$, localement bornée sur $\mathbb{R}_+$ et $O(1)$ lorsque $x \to 0^+$, on a

$$\mathrm{E}[C(Y_n)] \sim C(n). \tag{1.3}$$

Démonstration. Par hypothèse, on a lorsque $n \to \infty$ que

$$\int_n^\infty t^\delta f_n(t)dt = O(n^\delta) \qquad \text{et} \qquad \int_0^n t^{-\delta} f_n(t)dt = O(n^{-\delta}).$$

Le résultat est démontré si l'on établit que $\lim_{n\to\infty} \int_0^\infty f_n(t)\left[\frac{C(t)}{C(n)} - 1\right]dt = 0$. Soit $\varepsilon \in (0,1)$ et $W_n = \sup_{t\le n}\{t^\delta C(t)\}$, cette dernière étant bien définie par hypothèse sur C, on a

$$\begin{aligned}
\left|\int_0^{\varepsilon n} f_n(t)\left[\frac{C(t)}{C(n)} - 1\right]dt\right| &\le \int_0^{\varepsilon n} f_n(t)\left[\frac{C(t)}{C(n)} + 1\right]dt \\
&= \int_0^{\varepsilon n} \frac{f_n(t)C(t)}{C(n)}dt + \int_0^{\varepsilon n} f_n(t)dt \\
&\le \frac{W_{\varepsilon n}}{C(n)} \int_0^{\varepsilon n} t^{-\delta} f_n(t)dt + 1 \\
&\le \frac{W_{\varepsilon n}}{C(\varepsilon n)} \frac{C(\varepsilon n)}{C(n)} \int_0^n t^{-\delta} f_n(t)dt + 1.
\end{aligned}$$

À l'aide de la Définition 1.1 et du Théorème 1.3, il existe une constante $M_1 > 0$ indépendante de ε telle que

$$\limsup_{n\to\infty} \left|\int_0^{\varepsilon n} f_n(t)\left[\frac{C(t)}{C(n)} - 1\right]dt\right| \le \varepsilon^\delta M_1. \tag{1.4}$$

Similairement, si $W_n = \sup_{t\ge n}\{t^{-\delta}C(t)\}$,

$$\left|\int_{n/\varepsilon}^\infty f_n(t)\left[\frac{C(t)}{C(n)} - 1\right]dt\right| \le \frac{W_{n/\varepsilon}}{C(n)} \int_n^\infty f_n(t)t^\delta dt + 1$$

et donc pour $M_2 > 0$ indépendante de ε,

$$\limsup_{n\to\infty} \left| \int_{n/\varepsilon}^{\infty} f_n(t) \left[\frac{C(t)}{C(n)} - 1 \right] dt \right| \leq \varepsilon^{\delta} M_2. \tag{1.5}$$

Finalement, on a

$$\left| \int_{\varepsilon n}^{n/\varepsilon} f_n(t) \left[\frac{C(t)}{C(n)} - 1 \right] dt \right| \leq \sup_{t\in[\varepsilon n, n/\varepsilon]} \left| \frac{C(t)}{C(n)} - 1 \right| = \sup_{t\in[\varepsilon, 1/\varepsilon]} \left| \frac{C(tn)}{C(n)} - 1 \right| \to 0, \tag{1.6}$$

où on a utilisé le Théorème 1.2. En combinant les inégalités (1.4), (1.5) et (1.6) et en prenant $\varepsilon \to 0^+$, on peut conclure car

$$0 \leq \limsup_{n\to\infty} \left| \int_{0}^{\infty} f_n(t) \left[\frac{C(t)}{C(n)} - 1 \right] dt \right| \leq (M_1 + M_2)\varepsilon^{\delta}.$$

□

1.2 Théorie des valeurs extrêmes

La théorie des valeurs extrêmes est essentielle pour comprendre les travaux du Chapitre 2. Afin de bien comprendre cette théorie dans le cadre des mélanges Poisson, on présente les résultats pour des variables aléatoires positives. Les références principalement présentées dans cette section sont Resnick [1987], Coles [2001] et Leadbetter *et al.* [2012].

1.2.1 Formulation de la théorie

On suppose un échantillon $Y_1, \ldots, Y_n$ de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) d'une fonction de répartition quelconque F. En notant $\overline{Y}_n$ la moyenne empirique, le théorème limite central stipule que pour F à variance finie, $\sqrt{n}\left(\overline{Y}_n - \mathrm{E}[Y]\right)/\sqrt{\mathrm{Var}[Y]}$ converge en loi vers une normale centrée et réduite. D'une manière équivalente, le théorème limite central stipule qu'il existe des suites normalisantes a_n

et b_n telles que $(\overline{Y}_n - b_n)/a_n$ converge en loi vers une normale centrée et réduite. La théorie des valeurs extrêmes se fonde sur une question similaire où l'on s'intéresse plutôt au maximum de l'échantillon noté $M_n := \max(Y_1, \ldots, Y_n)$. Notons par $x_F = \sup\{x > 0 : F(x) < 1\}$ le point terminal de F, c'est-à-dire $F(x) = 1$ pour tout $x \geq x_F$. Selon la distribution, il est possible que x_F soit fini ou non. Puisque $\mathbb{P}(M_n \leq x) = F^n(x) < 1$ pour tout $x < x_F$, alors la distribution de M_n converge en distribution vers la loi dégénérée à x_F lorsque $n \to \infty$. En particulier, on peut démontrer que M_n converge presque sûrement vers x_F lorsque $n \to \infty$ [Leadbetter *et al.*, 2012, Corollaire 1.5.2]. Afin d'étudier les propriétés extrêmes de F, on a besoin de suites normalisantes $a_n \in \mathbb{R}_+$ et $b_n \in \mathbb{R}$ telles que

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \tag{1.7}$$

pour G une distribution non-dégénérée. Si ces suites existent, on dit alors que F ou Y est dans le *domaine d'attraction* de G. Fisher et Tippett [1928] démontrent qu'il existe trois domaines d'attraction possibles : Weibull, Gumbel et Fréchet. Gnedenko [1943] complète leurs résultats en démontrant que seulement ces trois domaines existent.

Théorème 1.6 (Fisher et Tippett [1928]; Gnedenko [1943])**.** Soient des suites $a_n \in \mathbb{R}_+$ et $b_n \in \mathbb{R}$ telles que $\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$ pour des distributions F et G non-dégénérées et tout x. Alors G est donnée par la distribution des valeurs extrêmes généralisées définie par

$$G_\gamma(x) = \begin{cases} \exp\left[-(1+\gamma x)^{-1/\gamma}\right] & \text{pour tout } x \text{ tel que } 1 + \gamma x > 0 \text{ si } \gamma \neq 0\,; \\ \exp\left[-e^{-x}\right] & \text{pour tout } x \in \mathbb{R} \text{ si } \gamma = 0. \end{cases} \tag{1.8}$$

Les trois domaines d'attraction possibles sont déterminés par le signe de γ du Théorème 1.6. Précisément, on a les domaines de Weibull, Gumbel et Fréchet pour $\gamma < 0$, $\gamma = 0$ et $\gamma > 0$ respectivement et on les note par $\mathcal{D}_-$, $\mathcal{D}_0$ et $\mathcal{D}_+$. Chacun de ces domaines possède différentes propriétés et caractérisations, notamment en lien avec la fonction de survie. En effet, la limite (1.7) peut être reformulée grâce au Théorème 1.5.1 de Leadbetter *et al.* [2012].

Théorème 1.7 (Leadbetter *et al.* [2012], p. 13)**.** La limite (1.7) est satisfaite si et seulement si les suites $a_n \in \mathbb{R}_+$ et $b_n \in \mathbb{R}$ sont telles que pour tout x

$$\lim_{n\to\infty} n\overline{F}(a_n x + b_n) = -\log G(x).$$

Ainsi, chaque domaine d'attraction peut être caractérisé par le comportement asymptotique de $\overline{F}$. Dans la prochaine sous-section, on explore chacune de ces caractérisations.

1.2.2 Domaines d'attraction

Domaine Fréchet

D'après le Théorème 1.6, les distributions F dans le domaine d'attraction de Fréchet possèdent des suites $a_n > 0$ et $b_n \in \mathbb{R}$ telles que (1.8) est satisfait pour $\gamma > 0$. En posant $\alpha = 1/\gamma$ et en remplaçant $(1 + \gamma x)$ par x, la limite est équivalente à

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \begin{cases} \exp[-x^{-\alpha}] & \text{si } x \geq 0 \\ 0 & \text{si } x < 0. \end{cases} \tag{1.9}$$

Ces distributions sont caractérisées par une queue lourde, c'est-à-dire leurs fonctions de survie possèdent un comportement à variation régulière [Resnick, 2007]. En notant par $\tau_n = \inf\{x > 0; F(x) \geq 1 - 1/n\}$, Gnedenko [1943] démontre les conditions nécessaires et suffisantes suivantes pour la limite (1.9).

Théorème 1.8 (Gnedenko [1943])**.** Une distribution F est dans $\mathcal{D}_+$ si et seulement si $x_F = \infty$ et $\overline{F} \in \mathcal{RV}_{-\alpha}$ pour $\alpha > 0$. Dans ce cas, la limite (1.9) est possible avec les suites $a_n = \tau_n$ et $b_n = 0$.

Il est donc impossible pour une distribution à queue finie d'être dans le domaine d'attraction $\mathcal{D}_+$. De ce théorème, on a également que F est toujours à queue lourde. De plus,

pour $r > \alpha$ et par l'inégalité de Markov, on a lorsque $x \to \infty$

$$\mathrm{E}[Y^r] \geq x^r \overline{F}(x) \sim C(x) x^{r-\alpha},$$

où $C \in \mathcal{RV}_0$. En vertu du Théorème 1.3, $\lim_{x\to\infty} x^r \overline{F}(x) = \infty$ et donc $\mathrm{E}[Y^r] = \infty$ pour tout $r > \alpha$. En fait, Galambos [1987] démontre que le r-ième moment pour $F \in \mathcal{D}_+$ existe seulement pour $r \in (0, \alpha)$.

Théorème 1.9 (Galambos [1987])**.** Soit $Y \sim F$ telle que $\overline{F} \in \mathcal{RV}_{-\alpha}$, alors $\mathrm{E}[Y^r] < \infty$ si $r \in (0, \alpha)$.

Domaine Weibull

Pour le cas des distributions F dans le domaine d'attraction de Weibull, il existe des suites $a_n > 0$ et $b_n \in \mathbb{R}$ telles que la distribution des maxima converge en loi vers la distribution (1.8) où $\gamma < 0$. En posant $\alpha = -1/\gamma$ et en remplaçant $(1 + \gamma x)$ par $-x$, cette limite est donnée par

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \begin{cases} \exp[-(-x)^\alpha] & \text{si } x < 0 \\ 1 & \text{si } x \geq 0. \end{cases} \tag{1.10}$$

Contrairement au domaine Fréchet, le domaine d'attraction de Weibull est caractérisé par des distributions à queue finie. La notion de variation régulière reste cependant toujours importante pour obtenir des conditions nécessaires et suffisantes.

Théorème 1.10 (Gnedenko [1943])**.** Une distribution F est dans $\mathcal{D}_-$ si et seulement si $x_F < \infty$ et $\overline{F}(x_F - 1/\bullet) \in \mathcal{RV}_{-\alpha}$. Dans ce cas, la limite (1.10) est possible avec les suites $a_n = x_F - \tau_n$ et $b_n = x_F$.

Évidemment, puisque les distributions dans ce domaine ont nécessairement une queue finie, tous les moments sont bien définis. Il est important de noter cependant qu'une distribution à queue finie n'est pas nécessairement dans le domaine de Weibull. Comme

on verra prochainement, certaines distributions à queue finie sont dans le domaine d'attraction de Gumbel. Finalement, lorsque $x_F > 0$, on note que $\overline{F}(x_F - 1/\bullet) \in \mathcal{RV}_{-\alpha}$ est équivalent à $\overline{F}\left(x_F \frac{\cdot}{\cdot+1}\right) \in \mathcal{RV}_{-\alpha}$ grâce à la Proposition 0.8 (iii) de Resnick [1987] puisque $x_F - \frac{1}{x} \sim x_F \frac{x}{x+1}$ lorsque $x \to \infty$. Ainsi, on peut adapter les conditions nécessaires et suffisantes de la façon suivante.

Théorème 1.11. Une distribution F avec $x_F > 0$ est dans $\mathcal{D}_-$ si et seulement si $x_F < \infty$ et $\overline{F}\left(x_F \frac{\cdot}{\cdot+1}\right) \in \mathcal{RV}_{-\alpha}$ pour $x > 0$, $a_n = x_F - \tau_n$ et $b_n = x_F$.

Domaine Gumbel

Finalement, les distributions F dans le domaine d'attraction de Gumbel sont telles qu'il existe des suites $a_n > 0$ et $b_n \in \mathbb{R}$ où pour tout $x \in \mathbb{R}$,

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp\left[-e^{-x}\right]. \tag{1.11}$$

Alors qu'il était possible de caractériser les domaines d'attraction de Fréchet et Weibull par les fonctions à variation régulière, ce n'est plus le cas pour le domaine de Gumbel. En effet, on a besoin d'une généralisation de cette propriété nommée Γ-variation. Cette propriété proposée par Haan [1970] n'est pas présentée dans cette thèse. Pour obtenir plus d'informations, voir Bingham *et al.* [1987] ou Resnick [1987]. La Γ-variation permet d'obtenir la caractérisation suivante obtenue également par Gnedenko [1943] quelques décennies avant Haan [1970].

Théorème 1.12 (Gnedenko [1943])**.** Une distribution F avec point terminal $x_F \in \mathbb{R}_+ \cup \{\infty\}$ est dans $\mathcal{D}_0$ si et seulement s'il existe une fonction strictement positive g, nommée fonction auxiliaire, telle que pour tout $x \in \mathbb{R}$,

$$\lim_{t\to x_F} \frac{1 - F\left(t + xg(t)\right)}{1 - F\left(t\right)} = e^{-x}.$$

Dans ce cas, la limite (1.11) est possible avec les suites $a_n = g(\tau_n)$ et $b_n = \tau_n$.

Pour les domaines précédents, chaque fonction de survie possédait une représentation de Karamata. Les fonctions de survie dans le domaine de Gumbel possèdent des représentations différentes dues à la Γ-variation [Balkema et Haan, 1972].

Théorème 1.13 (Balkema et Haan [1972])**.** Une distribution F avec point terminal $x_F \in \mathbb{R}_+ \cup \{\infty\}$ est dans $\mathcal{D}_0$ si et seulement s'il existe un $z_0 < x_F$ tel que pour tout $x \in (z_0, x_F)$ et $c > 0$

$$1 - F(x) = c(x) \exp\left[- \int_{z_0}^{x} \frac{dt}{g(t)} \right],$$

où $g(x) > 0$ est la fonction auxiliaire avec dérivée $g'(x)$ telle que $\lim_{x \to x_F} g'(x) = 0$ et $\lim_{x \to x_F} c(x) = c$.

Comme présenté dans les deux derniers résultats, il est possible d'avoir une distribution à queue finie ou infinie dans ce domaine d'attraction. En fait, il est possible de transformer toute variable aléatoire dans $\mathcal{D}_0$ et à valeurs dans $\mathbb{R}_+$ en une variable aléatoire à support fini dont le domaine est préservé.

Théorème 1.14 (Resnick [1987], p. 53)**.** Soient $T : \mathbb{R}_+ \to [a, b]$ une transformation strictement croissante deux fois différentiable telle que $\lim_{x \to \infty} xT''(x)/T'(x)$ est finie et Y une variable aléatoire avec distribution $F \in \mathcal{D}_0$. Alors $T(Y) \in \mathcal{D}_0$ avec comme suites normalisantes $\tilde{a}_n = T'(b_n)a_n$ et $\tilde{b}_n = T(b_n)$, où a_n et b_n sont les suites normalisantes de F.

Démonstration. Notons la transformation inverse de T par $I(x) = T^{-1}(x)$. Celle-ci existe et est strictement croissante par hypothèse. Soit $X := T(Y)$, alors $F_X(x) = F(I(x))$ pour $x \in [a, b]$. À l'aide du Théorème 1.13 et en posant $\tilde{c}(x) = c(I(x)) \to c > 0$ lorsque $x \to b$, on a

$$\begin{aligned} 1 - F_X(x) &= \tilde{c}(x) \exp\left[- \int_{z_0}^{I(x)} \frac{dt}{g(t)} \right] \\ &= \tilde{c}(x) \exp\left[- \int_{T(z_0)}^{x} \frac{dt}{T'(I(t))g(I(t))} \right]. \end{aligned}$$

Afin que $F_X(x)$ satisfasse la représentation du Théorème 1.13, il suffit de démontrer que $\tilde{g}(x) = T'(I(x))g(I(x))$ est positive et sa dérivée tend vers 0 lorsque $x \to b$. Par hypothèse, $\tilde{g}(x)$ est bien positive. Pour ce qui est de la dérivée, on a

$$\begin{aligned}\frac{d}{dx}\tilde{g}(x) &= \frac{T''(I(x))g(I(x))}{T'(I(x))} \\ &= \left[\frac{T''(I(x))I(x)}{T'(I(x))}\right]\left[\frac{g(I(x))}{I(x)}\right].\end{aligned}$$

Puisque $I(x) \to \infty$ lorsque $x \to b$ et $g(I(x))/I(x) \to 0$ [Resnick, 1987], on a bien que la limite converge vers 0 à l'aide de l'hypothèse sur le premier ratio. On conclut en calculant les constantes de normalisation. Soit a_n et b_n les suites normalisantes de F. On a par le Théorème 1.12 que

$$\begin{aligned}\tilde{b}_n &:= \inf\{x : F_X(x) \geq 1 - 1/n\} \\ &= \inf\{T(x) : F(x) \geq 1 - 1/n\} \\ &= T\left(\inf\{x : F(x) \geq 1 - 1/n\}\right) = T(b_n),\end{aligned}$$

et $\tilde{a}_n := \tilde{g}(\tilde{b}_n) = T'(b_n)g(b_n) = T'(b_n)a_n$. □

Un exemple de transformation qui satisfait les conditions du Théorème 1.14 est $T(x) = bx/(x+1) + a$. En particulier, si $a = 0$, $b = 1$ et $Y \sim \text{Exp}(1)$, la distribution exponentielle, alors $T(Y)$ est une distribution avec support $[0, 1]$ dans $\mathcal{D}_0$. En effet, on peut facilement montrer que $Y \in \mathcal{D}_0$ avec $a_n = 1$ et $b_n = \log n$. Ainsi, les suites normalisantes de $T(Y)$ sont données par $\tilde{a}_n = (1 + \log n)^{-2}$ et $\tilde{b}_n = \log n/(\log n + 1)$. On conclut cette section avec les moments de $F \in \mathcal{D}_0$. Si $x_F < \infty$, il est évident que $\mathrm{E}[Y^r] < \infty$ pour tout $r > 0$. Il s'avère que cette propriété est toujours satisfaite lorsque $x_F = \infty$.

Théorème 1.15 (Resnick [1987], p. 52)**.** Soit $Y \in \mathcal{D}_0$, alors $\mathrm{E}[Y^r] < \infty$ pour tout $r > 0$.

1.2.3 Cas discret

Jusqu'à présent, aucune hypothèse n'a été faite sur la nature de la distribution. En effet, seulement des hypothèses sur le comportement asymptotique de la fonction survie ont été établies. Dans le cas de distributions continues, il est généralement possible d'identifier le domaine d'attraction. Il existe cependant quelques exceptions. Par exemple, Subramanya [1994] démontre que la log-Pareto ne possède aucune domaine d'attraction. Pour les lois discrètes, ces exceptions sont plus fréquentes. En effet, plusieurs distributions discrètes ne possèdent aucune suite normalisante pour M_n. Pour bien comprendre cette difficulté, on doit reformuler le Théorème 1.7 en termes de suites u_n quelconques. À l'aide d'une preuve similaire, on obtient que $\lim_{n\to\infty} n\overline{F}(u_n) = \tau$ si et seulement si $\lim_{n\to\infty} F^n(u_n) = e^{-\tau}$ pour $\tau \in \mathbb{R}_+$. On retrouve bien le Théorème 1.7 en posant $u_n = a_n x + b_n$ et $\tau = -\log G(x)$. Or dans ce cadre général, la convergence de M_n peut être caractérisée [Leadbetter *et al.*, 2012].

Théorème 1.16 (Leadbetter *et al.* [2012], p. 24)**.** Pour une distribution F avec point terminal x_F, il existe une suite u_n telle que $\lim_{n\to\infty} F^n(u_n) = \tau \in \mathbb{R}_+$ si et seulement si

$$\lim_{x\to x_F} \frac{\overline{F}(x)}{\overline{F}(x-)} = 1,$$

où $x-$ la limite à gauche de x.

En particulier, si F est une distribution discrète avec comme support un sous-ensemble de $\mathbb{N}$, alors il existe une suite u_n telle que la distribution de M_n converge si et seulement si

$$\lim_{n\to x_F} \frac{\overline{F}(n)}{\overline{F}(n-1)} = 1.$$

Il s'avère que cette contrainte peut être difficile à satisfaire pour plusieurs distributions bien connues. Directement, on peut conclure par le Théorème 1.16 qu'aucune distribution

discrète à support fini ne possède de domaine d'attraction puisque $\overline{F}(x_F)/\overline{F}(x_F - 1) = 0$. Ainsi, le domaine de Weibull est désuet dans le cadre des distributions discrètes. De même pour le domaine de Gumbel avec $x_F < \infty$. Pour ce qui est des distributions où $x_F = \infty$, considérons la loi géométrique (Géo) comme exemple. Soit $p \in (0, 1)$, la fonction de survie de celle-ci est donnée par $\overline{F}(n) = p^{n+1}$ pour $n \in \mathbb{N}$. En vertu du Théorème 1.16, on peut conclure qu'aucune suite normalisante de M_n n'existe puisque pour tout n,

$$\frac{\overline{F}(n)}{\overline{F}(n-1)} = p < 1.$$

La loi géométrique peut être vue comme une discrétisation de la loi exponentielle, c'est-à-dire $\lfloor X \rfloor \sim \text{Géo}(p)$ lorsque $X \sim \text{Exp}(-\log p)$. Cependant, la loi exponentielle est bien dans le domaine d'attraction de Gumbel. On peut donc conclure que discrétiser une loi continue ne préserve pas nécessairement le domaine d'attraction. Shimura [2012] étudie plus en détail ce phénomène. Lorsque $u_n = a_n x + b_n$, Anderson [1970] démontre que la propriété de *queue longue* est nécessaire afin qu'une loi discrète F soit élément d'un domaine d'attraction.

Théorème 1.17 (Anderson [1970])**.** Soit F une distribution discrète avec support $\mathbb{N}$. Si F possède un domaine d'attraction, alors elle est à queue longue, c'est-à-dire

$$\lim_{n\to\infty} \frac{\overline{F}(n+1)}{\overline{F}(n)} = 1.$$

Comme au Théorème 1.16, cette condition exclut plusieurs distributions discrètes bien connues. Par exemple, la loi de Poisson ($\mathcal{P}$), la loi binomiale négative ($\mathcal{NB}$) ou la loi géométrique sont toutes des distributions sans queue longue. Malgré l'absence d'un domaine d'attraction, il est possible de bien approcher les extrêmes de certaines distributions discrètes à l'aide de la loi de Gumbel. En effet, Anderson [1970] démontre, pour F une

distribution discrète et $\alpha > 0$, qu'il existe une suite b_n telle que

$$\begin{aligned} \limsup_{n\to\infty} F^n(x + b_n) &\leq \exp\left(-e^{-\alpha x}\right) \\ \liminf_{n\to\infty} F^n(x + b_n) &\geq \exp\left(-e^{-\alpha(x-1)}\right) \end{aligned} \tag{1.12}$$

si et seulement si

$$\lim_{n\to\infty} \frac{\overline{F}(n+1)}{\overline{F}(n)} = e^{-\alpha} \in (0, 1). \tag{1.13}$$

Par conséquent, les limites supérieure et inférieure sont toutes deux bornées par des fonctions de répartition Gumbel. Les distributions binomiale négative et géométrique sont des exemples possédant cette propriété. En particulier, on peut démontrer que la loi géométrique de paramètre p satisfait la condition (1.13) en posant $\alpha = -\log p$ et $b_n = \log(n)/\alpha - 1$ dans (1.12). Ceci est illustré avec la Figure 1.1 pour $p = 0.7$ et $n = 1000$. De plus Shimura [2012] démontre que toute loi satisfaisant la limite (1.13) provient d'une loi continue dans le domaine $\mathcal{D}_0$ qui a été discrétisée. Comme on l'a remarqué, cela concorde avec la relation entre la loi géométrique et exponentielle. Dans ces deux cas, on dit que la loi F est *proche du domaine de Gumbel* et le degré de proximité est déterminé par la valeur de α. En effet, la limite (1.13) s'approche de 1, c'est-à-dire F s'approche d'une queue longue, si α s'approche de 0. Inversement, F s'éloigne du domaine de Gumbel si $\alpha \to \infty$. Anderson [1970] étudie également cette dernière situation. Il démontre qu'il existe une suite d'entiers I_n telle que pour une distribution F discrète,

$$\lim_{n\to\infty} \mathbb{P}(M_n = I_n \text{ ou } I_n + 1) = 1$$

si et seulement si

$$\lim_{n\to\infty} \frac{\overline{F}(n+1)}{\overline{F}(n)} = 0. \tag{1.14}$$

Précisément, si une distribution F satisfait (1.14), ses maxima alternent presque sûrement entre deux valeurs entières. Autrement dit, la distribution F est *éloignée de tout domaine d'attraction*. On peut facilement montrer que la loi de Poisson satisfait à (1.14). Ceci est illustré à la Figure 1.2 pour la distribution de M_n, $\lambda = 5$ et $n \in \{10, 10^2, 10^3, 10^4\}$.

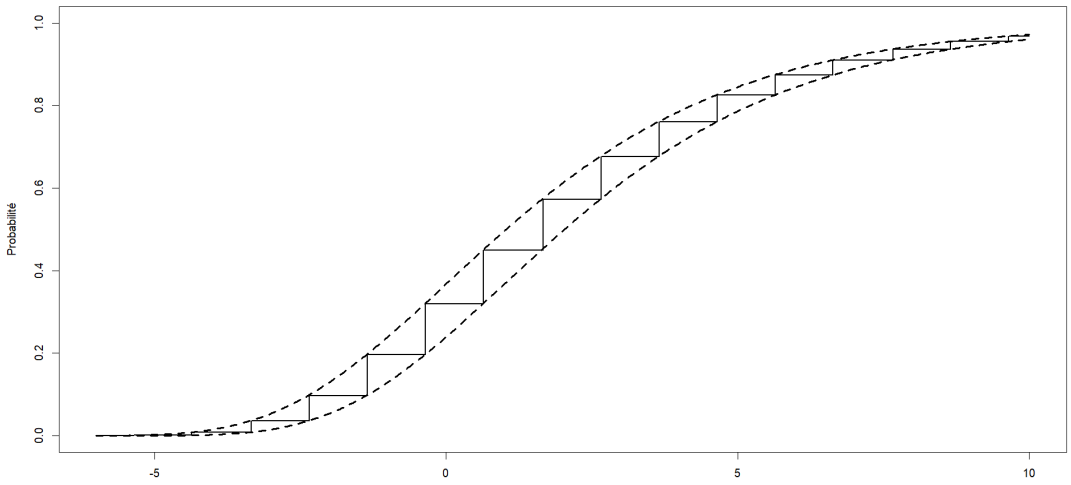


FIGURE 1.1 – Distribution de M_n normalisé pour la loi Géo(0.7), $n = 1000$ et les bornes (1.12) avec $\alpha = -\log p$.

1.3 Mélanges Poisson

La distribution de Poisson apparaît naturellement lors de la modélisation de données de comptage. Une variable aléatoire Y est distribuée selon une loi Poisson de paramètre $\lambda \in \mathbb{R}_+$ si sa fonction de masse est donnée par

$$\mathbb{P}(Y = n) = \frac{\lambda^n}{n!} e^{-\lambda},\ n \in \mathbb{N}.$$

On note celle-ci par $Y \sim \mathcal{P}(\lambda)$. Une propriété bien connue de cette distribution est l'égalité $\mathrm{Var}[Y] = \mathrm{E}[Y]$. Ainsi, on dit que la loi de Poisson possède une *dispersion nulle*. Cependant, les données de comptage sont généralement *surdispersées*, c'est-à-dire $\mathrm{Var}[Y] > \mathrm{E}[Y]$. Une solution naturelle à cet inconvénient est de supposer λ aléatoire. On dit alors que Y est un *mélange Poisson* si $Y|\lambda \sim \mathcal{P}(\lambda)$ et $\lambda \sim F$, une distribution dans $\mathbb{R}_+$. On note respectivement par F_M, $\overline{F}_M$ et P_M la fonction cumulative, fonction de survie et la fonction de masse du mélange résultant. Dans cette section, on présente les propriétés des mélanges Poisson essentielles à cette thèse. Pour en savoir plus sur ces distributions, Karlis et Xekalaki [2005] offrent une revue détaillée.

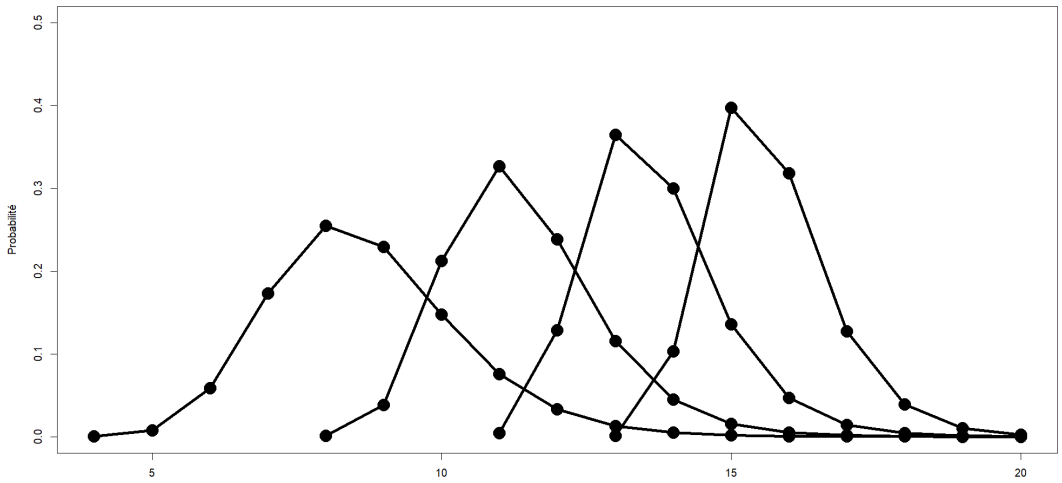


FIGURE 1.2 – Distribution de M_n pour la loi $\mathcal{P}(5)$ et $n \in \{10, 10^2, 10^3, 10^4\}$.

Par définition, la fonction de masse du mélange Poisson est donnée par

$$P_M(n) = \int_\Lambda \frac{\lambda^n}{n!} e^{-\lambda} dF(\lambda),$$

où $\Lambda \subseteq \mathbb{R}_+$ est le support de la loi de mélange F. Feller [1943] a démontré que les mélanges Poisson sont identifiables, c'est-à-dire si $Y \sim F_M$ telle que $\int_\Lambda \frac{\lambda^n}{n!} e^{-\lambda} dF_1(\lambda) = \int_\Lambda \frac{\lambda^n}{n!} e^{-\lambda} dF_2(\lambda)$, alors $F_1(\lambda) = F_2(\lambda)$. Ainsi, chaque mélange Poisson est caractérisé par leur distribution sur λ. Par souci de simplicité, on suppose pour le reste de cette section que $\Lambda = (x_0, x_1)$ où $0 \leq x_0 < x_1 \leq \infty$. Notez cependant qu'il est possible que Λ soit une collection finie d'intervalles disjoints menant ainsi à d'autres distributions distinctes. Puisqu'on suppose que F n'est pas une distribution dégénérée, les mélanges Poisson sont bien surdispersés car

$$\mathrm{Var}[Y] = \mathrm{Var}[\lambda] + \mathrm{E}[\lambda] > \mathrm{E}[\lambda] = \mathrm{E}[Y].$$

De plus, les mélanges Poisson sont préférables lorsque les données présentent un excès de zéros, mais également lorsque celles-ci possèdent de grandes valeurs. En effet, Feller [1943] démontre le résultat suivant.

Théorème 1.18 (Feller [1943])**.** Soit $\lambda \sim F$ avec $\mu := \mathrm{E}[\lambda] \in \mathbb{R}_+$ et $Y \mid \lambda \sim \mathcal{P}(\lambda)$, alors

$$\mathbb{P}(Y = 0) \geq \mathbb{P}(Y = 0 \mid \lambda = \mu).$$

Shaked [1980] généralise ce résultat et démontre que les mélanges Poisson possèdent une queue plus lourde. Précisément, il démontre le résultat suivant.

Théorème 1.19 (Shaked [1980])**.** Soit $\lambda \sim F$ avec $\mu := \mathrm{E}[\lambda] \in \mathbb{R}_+$ et $Y \mid \lambda \sim \mathcal{P}(\lambda)$, alors $\mathbb{P}(Y = n) - \mathbb{P}(Y = n \mid \lambda = \mu)$ possède les changements de signe suivants lorsque n varie de 0 à l'infini : positif, négatif et positif.

En combinant ces deux théorèmes, P_M permet de modéliser des variables aléatoires avec un plus grand excès de zéros, mais également des valeurs extrêmes plus probables comparativement à la loi de Poisson. Afin d'étudier le comportement extrême des mélanges Poisson, on a besoin d'une représentation de $\overline{F}_M$ qui utilise la fonction de survie $\overline{F}$ de λ.

Théorème 1.20 (Karlis et Xekalaki [2005])**.** Soit $Y \mid \lambda \sim \mathcal{P}(\lambda)$ et $\lambda \sim F$ avec $F(x_1) - F(x_0) = 1$, alors

$$\overline{F}_M(n) := \mathbb{P}(Y > n) = \mathbb{P}(Y > n \mid \lambda = x_0) + \int_{x_0}^{x_1} \frac{\lambda^n e^{-\lambda}}{n!}(1 - F(\lambda))d\lambda.$$

Démonstration. À l'aide de l'intégration par parties, on a

$$\begin{aligned}
\mathbb{P}(Y > n) &= \sum_{k=n+1}^{\infty} \int_{x_0}^{x_1} \frac{\lambda^k e^{-\lambda}}{k!} dF(\lambda) \\
&= \sum_{k=n+1}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!}(F(\lambda) - 1)\Big|_{x_0}^{x_1} + \int_{x_0}^{x_1} \sum_{k=n+1}^{\infty} \left[\frac{\lambda^{k-1}}{(k-1)!} - \frac{\lambda^k}{k!}\right](1 - F(\lambda))d\lambda \\
&= \mathbb{P}(Y > n \mid \lambda = x_0) + \int_{x_0}^{x_1} \frac{\lambda^n e^{-\lambda}}{n!}(1 - F(\lambda))d\lambda,
\end{aligned}$$

où on a utilisé le fait que la somme à l'intérieur de l'intégrale est télescopique. □

En particulier, lorsque $x_0 = 0$, le premier terme du Théorème 1.20 disparaît. En fait, ce terme peut être ignoré lorsque $n \to \infty$, et ce peu importe la valeur de x_0. Pour cette thèse, il suffit d'étudier les cas où $\lambda \sim F$ tels que 1) $\mathrm{E}[\lambda] < \infty$, ou 2) $\overline{F} \in \mathcal{RV}_{-\alpha}$. On obtient les corollaires suivants.

Corollaire 1.1. Soit $Y \mid \lambda \sim \mathcal{P}(\lambda)$ et $\lambda \sim F$ tel que $\mu := \mathrm{E}[\lambda] < \infty$. Lorsque $n \to \infty$,

$$\mathbb{P}(Y > n) \sim \int_{x_0}^{x_1} \frac{\lambda^n e^{-\lambda}}{n!}(1 - F(\lambda))d\lambda.$$

Démonstration. On remarque $\mu > x_0$, autrement $\lambda = x_0$ presque sûrement. Pour obtenir le résultat, il suffit de démontrer que $\mathbb{P}(Y > n \mid \lambda = x_0) = o(\mathbb{P}(Y > n))$ lorsque $n \to \infty$. Puisque les fonctions de survie de Y et $Y \mid \lambda$ forment des suites strictement décroissantes, on applique le théorème de Stolz-Cesàro pour évaluer la limite. Précisément, si on a

$$\lim_{n\to\infty} \frac{\mathbb{P}(Y > n+1 \mid \lambda = x_0) - \mathbb{P}(Y > n \mid \lambda = x_0)}{\mathbb{P}(Y > n+1) - \mathbb{P}(Y > n)} = \lim_{n\to\infty} \frac{\mathbb{P}(Y = n+1 \mid \lambda = x_0)}{\mathbb{P}(Y = n+1)} = 0,$$

alors la limite qui nous intéresse converge également vers 0. Par le Théorème 1.19, $P_M(n) \geq \mathbb{P}(Y = n \mid \lambda = \mu)$ pour n suffisamment grand. On peut donc conclure puisque

$$\lim_{n\to\infty} \frac{\mathbb{P}(Y = n+1 \mid \lambda = x_0)}{\mathbb{P}(Y = n+1)} \leq \lim_{n\to\infty} \frac{\mathbb{P}(Y = n+1 \mid \lambda = x_0)}{\mathbb{P}(Y = n+1 \mid \lambda = \mu)} = \lim_{n\to\infty} \left(\frac{x_0}{\mu}\right)^{n+1} e^{\mu - x_0} = 0.$$

□

Corollaire 1.2. Soit $Y \mid \lambda \sim \mathcal{P}(\lambda)$ et $\lambda \sim F$ tel que $\overline{F} \in \mathcal{RV}_{-\alpha}$. Lorsque $n \to \infty$,

$$\mathbb{P}(Y > n) \sim \int_{x_0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!}(1 - F(\lambda))d\lambda.$$

Démonstration. Comme on démontre au Chapitre 2, on a par Théorème 1.5

$$\int_{x_0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!}(1 - F(\lambda))d\lambda = \int_0^{\infty} \frac{\lambda^n e^{-\lambda}}{n!}(C(\lambda)\lambda^{-\alpha}\mathbb{1}_{\{\lambda > x_0\}})d\lambda \sim C(n)n^{-\alpha}$$

pour $C \in \mathcal{RV}_0$. Puisque la fonction de survie de la loi de Poisson est telle que

$$\mathbb{P}(Y > n \mid \lambda = x_0) = \frac{\gamma(n+1, x_0)}{\Gamma(n+1)} \sim \frac{x_0^{n+1} e^{x_0}}{\Gamma(n+1)},$$

où $\gamma(n, x_0)$ est la fonction gamma incomplète [Olver *et al.*, 2010], alors par l'approximation de Stirling, c'est-à-dire pour $c > 0$, $\Gamma(x + c) \sim \sqrt{2\pi}e^{-x}x^{x+c-1/2}$ lorsque $x \to \infty$, on obtient

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathbb{P}(Y > n \mid \lambda = x_0)}{\int_{x_0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!}(1 - F(\lambda))d\lambda} &= \lim_{n\to\infty} \frac{x_0^{n+1}e^{-x_0}}{\Gamma(n+2)C(n)n^{-\alpha}} \\ &= x_0 e^{-x_0} \lim_{n\to\infty} \left(\frac{x_0^n}{\Gamma(n-\alpha)}\right)\left(\frac{1}{C(n)n^2}\right) = 0.\end{aligned}$$

□

En combinant ces divers résultats avec ceux de la théorie des valeurs extrêmes présentés à la section 1.2, nous présenterons au Chapitre 2 une analyse détaillée des comportements extrêmes des mélanges Poisson.

1.4 Modèle multivarié discret Splitting

Comme dans le cas univarié, l'utilisation de la loi de Poisson semble naturelle pour modéliser les données de comptage multivariées. Une première approche proposée par Campbell [1934] et généralisée par Teicher [1954] est l'utilisation de $J+1$ variables aléatoires Poisson indépendantes afin de créer une distribution multivariée de dimension J avec marginales Poisson. Plus précisément, ils supposent que $(Y_j)_{j=0}^{J}$ sont des variables Poisson indépendantes et chaque marginale est définie par $Y_j + Y_0$ pour $j \in \{1, \ldots, J\}$. Puisque la somme de variables aléatoires indépendantes Poisson reste Poisson, on a bien que les marginales sont distribuées selon cette loi. Cependant, les corrélations de ce modèle sont nécessairement positives et chaque interaction est déterminée par Y_0. Une modification possible, proposée par Schulz *et al.* [2021], est l'introduction de variables Poisson co-monotones. Les interactions dans ce cas sont plus intéressantes, mais les corrélations restent toujours

du même signe. Pour plus de détails concernant les modèles multivariés dérivés de la loi de Poisson, voir par exemple Inouye *et al.* [2017].

Une seconde approche peut être l'utilisation des copules avec des marginales discrètes. On rappelle qu'une copule de dimension J est une distribution $\mathcal{C} : [0,1]^J \to [0,1]$ satisfaisant

— $\mathcal{C}(u_1, \ldots, u_J) = 0$ lorsque $u_j = 0$ pour au moins un $j \in \{1, \ldots, J\}$;

— $\mathcal{C}(u_1, \ldots, u_J) = u_i$ lorsque $u_j = 1$ pour tout $j \in \{1, \ldots, J\}$ et $i \neq j$;

— Pour tout hypercube $[\mathbf{a}, \mathbf{b}] \subseteq [0,1]^J$, $\mathcal{C}([\mathbf{a}, \mathbf{b}]) \geq 0$.

Les travaux fondamentaux de Sklar [1959] démontrent que pour une distribution F de dimension J et marginales $F_1, \ldots, F_J$, il existe une copule $\mathcal{C}$ unique sur $\mathrm{Im}(F_1) \times \cdots \times \mathrm{Im}(F_J)$ telle que $F(\mathbf{y}) = \mathcal{C}(F_1(y_1), \ldots, F_J(y_J))$, où $\mathrm{Im}(F_j)$ est l'image de la marginale F_j. Inversement, pour une copule $\mathcal{C}$ et distributions F_j, alors $\mathcal{C}(F_1(x_1), \ldots, F_J(x_J))$ définit une distribution de dimension J avec marginales F_j. Dans le cadre des distributions continues, les copules décrivent uniquement la structure de dépendance entre les marginales. Entre autres, les copules permettent de séparer l'analyse multivariée en deux composantes : les marginales et la structure de dépendance. Toutefois, l'unicité de la copule n'est plus vérifiée lorsque les distributions marginales sont discrètes, ce qui a pour conséquence de rendre les modèles non identifiables [Genest et Nešlehová, 2007]. Notez qu'il est possible d'utiliser les copules dans ce contexte, voir aussi par exemple Panagiotelis *et al.* [2012]. Cependant, il est important d'être précautionneux lorsqu'on utilise les copules avec des distributions discrètes.

Peyhardi et Fernique [2017], Jones et Marchand [2019] et Peyhardi *et al.* [2021] proposent une démarche différente, mais aussi simple que les deux approches précédentes, nommée *modèle Sums and Shares* ou *modèle Splitting*. Ces deux modèles supposent que la somme des composantes du vecteur $\mathbf{Y}$ est distribuée selon une loi discrète univariée et que celle-ci est "divisée" aléatoirement parmi les J composantes de $\mathbf{Y}$. Pour cette thèse, nous allons privilégier la terminaison Splitting lorsque nous discutons de cette classe de

distributions. Dans cette section, on présente la définition du modèle Splitting et les propriétés nécessaires pour comprendre les travaux des Chapitres 3 et 4.

1.4.1 Définitions

Soit $\mathbf{Y} = (Y_1, \ldots, Y_J) \in \mathbb{N}^J$. Le modèle Splitting est défini en deux étapes. On suppose premièrement que $|\mathbf{Y}| := \sum_{j=1}^{J} Y_j$ est distribuée selon une loi $\mathcal{L}(\boldsymbol{\psi})$ dans $\mathbb{N}$ avec paramètre $\boldsymbol{\psi}$. Deuxièmement, on suppose $\mathbf{Y}$ conditionnellement à $|\mathbf{Y}| = n$ est distribuée selon une loi singulière $\mathcal{S}_{\Delta_n}(\boldsymbol{\theta})$ avec support le simplexe discret $\Delta_n := \{\mathbf{y} \in \mathbb{N}^J : |\mathbf{y}| = n\}$ et paramètre $\boldsymbol{\theta}$. Ainsi, la fonction de masse de $\mathbf{Y}$ est telle que

$$\mathbf{p}(\mathbf{y}) = \mathbf{p}_{|\mathbf{Y}|=n}(\mathbf{y})\mathbf{p}(|\mathbf{Y}| = n)$$

où $n = |\mathbf{y}|$ et $\mathbf{p}_{|\mathbf{Y}|=n}$ est la fonction de masse de $\mathcal{S}_{\Delta_n}$. Si le vecteur $\mathbf{Y}$ est distribué selon un tel modèle Splitting, on note celui-ci par

$$\mathbf{Y} \sim \mathcal{S}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi}),$$

où $\wedge$ est l'opérateur de mélange. Cette dernière notation a été introduite dans les travaux de Peyhardi *et al.* [2021]. On peut interpréter cette approche par la modélisation d'une ressource qui est divisée aléatoirement dans J catégories. Cette simple division permet d'obtenir une grande famille de lois multivariées discrètes. Jones et Marchand [2019] et Bhagwat [2019] étudient le cas particulier où $\mathcal{S}_{\Delta_n}$ est une Dirichlet-multinomiale et $\mathcal{L}$ une binomiale négative. Peyhardi *et al.* [2021] et Peyhardi [2023] généralisent leurs résultats en étudiant le cas où $\mathcal{S}_{\Delta_n}$ est Pólya ou quasi-Pólya [Janardan et Schaeffer, 1977] et $\mathcal{L}$ est une distribution quelconque. Précisément, la loi est une quasi-Pólya $\mathcal{QP}^{[c,d]}_{\Delta_n}(\boldsymbol{\theta})$ de paramètres $\boldsymbol{\theta}$, $c \in \{-1, 0, 1\}$ et $d \geq 0$ si sa fonction de masse est telle que pour $\mathbf{y} \in \Delta_n$,

$$\mathbf{p}_{|\mathbf{Y}|=n}(\mathbf{y}) = \frac{1}{a^{[c,d]}_{|\boldsymbol{\theta}|}(n)} \prod_{j=1}^{J} a^{[c,d]}_{\theta_j}(y_j), \tag{1.15}$$

où $a_\theta^{[c,d]}(n) = \frac{\theta(\theta+dn)_{(n,c)}}{(\theta+dn)n!}$ et $(x)_{(n,c)}$ est la factorielle généralisée

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

Comme cas particulier, on utilise le symbole de Pochhammer $(x)_n := (x)_{(n,1)}$ pour la factorielle ascendante. Similairement, la factorielle décroissante est définie par $(x)_{(n,-1)} = (-1)^n(-x)_n$. Lorsque $d = 0$, on retrouve les lois de Pólya, notées par $\mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta})$, avec fonctions de masse données par

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

Celles-ci possèdent trois formes différentes selon la valeur de c : l'hypergéométrique ($\mathcal{H}_{\Delta_n}$), la multinomiale ($\mathcal{M}_{\Delta_n}$) et la Dirichlet-multinomiale ($\mathcal{DM}_{\Delta_n}$). Elles correspondent respectivement aux valeurs $c = -1, c = 0$ et $c = 1$. Pour $d > 0$, il existe deux classes de lois possibles : la quasi-multinomiale ($\mathcal{QM}_{\Delta_n}$) pour $c = 0$ et la quasi-Dirichlet-multinomiale ($\mathcal{QDM}_{\Delta_n}$) pour $c = 1$. Toujours lorsque $d > 0$, il est possible de démontrer que la distribution quasi-Pólya où $c = -1$ est équivalente à la même loi lorsque $c = 1$. Ainsi, il n'est pas nécessaire de définir la quasi-hypergéométrique. En général, les composantes de $\boldsymbol{\theta}$ prennent des valeurs réelles positives, la seule exception étant la loi hypergéométrique où $\boldsymbol{\theta} \in \mathbb{N}_+^J$ est telle que $|\boldsymbol{\theta}| \geq n$ afin d'obtenir une distribution bien définie. Finalement, (1.15) est bien une fonction de masse. En effet, Janardan et Schaeffer [1977] démontrent que les fonctions $a_\theta^{[c,d]}$ possèdent la propriété de *convolution additive*, c'est-à-dire

$$a_{\theta+\gamma}^{[c,d]}(n) = \sum_{k=0}^{n} a_\theta^{[c,d]}(k) a_\gamma^{[c,d]}(n-k).$$

Par récurrence, on peut facilement démontrer que

$$a_{|\boldsymbol{\theta}|}^{[c,d]}(n) = \sum_{\mathbf{y} \in \Delta_n} \prod_{j=1}^{J} a_{\theta_j}^{[c,d]}(y_j).$$

1.4.2 Propriétés

À l'aide de la convolution additive, Jones et Marchand [2019], Peyhardi *et al.* [2021] et Peyhardi [2023] démontrent plusieurs propriétés essentielles aux contributions présentées dans le Chapitre 3. Pour cette thèse, on présente ces propriétés lorsque $\mathcal{S}_{\Delta_n}(\boldsymbol{\theta})$ est une distribution de Pólya. Dans ce cas, on dit que le vecteur aléatoire $\mathbf{Y}$ suit une distribution *Pólya Splitting* et est noté par

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

Notons que les démonstrations des résultats suivants utilisent la propriété de convolution additive. Ainsi, lorsque $\mathcal{S}_{\Delta_n}$ est quasi-Pólya, ces mêmes propriétés sont toujours satisfaites. De plus, remarquons pour $c = -1$ que la loi $\mathcal{L}$ doit être à support fini. En effet, puisque $|\boldsymbol{\theta}| \geq n$ pour l'hypergéométrique lorsque n est fixé, cette contrainte doit être satisfaite lorsque n est aléatoire. Le modèle Pólya Splitting est adéquat pour $c = -1$ si et seulement si le support de $\mathcal{L}$ est borné par une valeur entière m et $|\boldsymbol{\theta}| \geq m$. Finalement, on rappelle que la loi de Pólya est singulière, c'est-à-dire que son support possède $J - 1$ degrés de liberté. Par exemple si $J = 2$, $Y_2 = n - Y_1$ et donc la loi de Y_1 détermine le couple (Y_1, Y_2). Ainsi, la Pólya univariée est définie lorsque $J = 2$ et on note celle-ci par $\mathcal{P}^{[c]}_n(\theta, \tau)$.

Distributions marginales

L'une des propriétés importantes de cette classe de distributions est que toute marginale reste dans cette classe. Plus précisément, pour tout ensemble d'indices $\mathcal{J} \subseteq \{1, \ldots, J\}$, la marginale multivariée $\mathbf{Y}_{\mathcal{J}} \coloneqq (Y_j)_{j \in \mathcal{J}}$ est elle-même une distribution Pólya Splitting.

Théorème 1.21 (Peyhardi *et al.* [2021])**.** Soit $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ et $\mathcal{J} \subseteq \{1, \ldots, J\}$ alors,

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

où $\boldsymbol{\theta}_{\mathcal{J}} = (\theta_j)_{j\in\mathcal{J}}$ et $\boldsymbol{\theta}_{-\mathcal{J}} = (\theta_j)_{j\in\mathcal{J}^c}$.

En particulier, le Théorème 1.21 stipule que la somme $|\mathbf{Y}_{\mathcal{J}}|$ est distribuée selon la loi $\mathcal{P}_n^{[c]}(|\boldsymbol{\theta}_{\mathcal{J}}|, |\boldsymbol{\theta}_{-\mathcal{J}}|) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$. On note également que

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

lorsque $\mathcal{J}$ est un singleton.

Stabilité par composition

On peut interpréter la distribution de Y_j comme la réduction de $|\mathbf{Y}|$ dans la catégorie j. D'une certaine façon, la variable $|\mathbf{Y}|$ est "endommagée" par la Pólya. Rao [1965], Bol'shev [1965], Patil et Ratnaparkhi [1975] et Rao et Janardan [1984] ont initialement introduit cette idée d'endommagement comme technique de caractérisation de certaines distributions discrètes. Autrement, Joe [1996] et Peyhardi [2023] présentent cette composition comme un opérateur dit *Thinning*. Ce dernier est particulièrement utile pour définir des séries temporelles autorégressives dans le cadre des données de comptage. On réfère à Davis *et al.* [2021] pour plus de détails sur cette application. Si L est une loi telle que

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

où $\tilde{\boldsymbol{\psi}}$ est un vecteur de paramètres modifiés par la composition, alors on dit que $\mathcal{L}$ est *stable par composition* pour $\mathcal{P}_n^{[c]}(\boldsymbol{\theta})$. Peyhardi [2023] démontre qu'il existe trois familles de lois stables par composition : la loi de Pólya, la série entière additive et l'inverse Pólya. Pour la première famille, la propriété de convolution additive permet de démontrer que

$$\mathcal{P}_m^{[c]}(\theta, \tau + \lambda) = \mathcal{P}_n^{[c]}(\theta, \tau) \underset{n}{\wedge} \mathcal{P}_m^{[c]}(\theta + \tau, \lambda).$$

Pour ce qui est des distributions de type série entière additive, elles sont données par la binomiale ($\mathcal{B}_n$), la Poisson ($\mathcal{P}$) et la binomiale négative ($\mathcal{NB}$) pour $c = -1$, $c = 0$ et $c = 1$

respectivement (voir Tableau 1.1). Ce qui distingue cette famille des deux autres est la propriété d'indépendance entre les marginales. En effet, Bol'shev [1965] démontre pour $J = 2$ que ce sont seulement ces distributions qui induisent l'indépendance dans le modèle Pólya Splitting et Peyhardi [2023] généralise ce résultat pour une dimension quelconque.

Théorème 1.22 (Peyhardi [2023])**.** Soit $\mathbf{Y} \sim \mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, alors les marginales Y_j sont indépendantes si et seulement si $\mathcal{L}$ est donnée par

- $\mathcal{B}_{|\boldsymbol{\theta}|}(\pi)$ pour $c = -1$ et $\boldsymbol{\theta} \in \mathbb{N}_+^J$;
- $\mathcal{P}(\lambda)$ pour $c = 0$ et tout $\lambda \in \mathbb{R}_+$;
- $\mathcal{NB}(|\boldsymbol{\theta}|, p)$ pour $c = 1$, $\boldsymbol{\theta} \in \mathbb{R}_+^J$ et $p \in (0, 1)$.

Selon le Théorème 1.22, les distributions marginales sont toujours de type série entière additive. En particulier, la fonction de masse jointe de $\mathbf{Y}$ est simplement le produit des J densités marginales.

Distributions	Paramètres	Fonction de masse
$\mathcal{B}_n(\pi)$	$n \in \mathbb{N}$, $\pi \in (0,1)$	$\binom{n}{y}\pi^y(1-\pi)^{n-y}$
$\mathcal{P}(\lambda)$	$\lambda \in \mathbb{R}_+$	$\frac{\lambda^y}{y!}e^{-\lambda}$
$\mathcal{NB}(\alpha, p)$	$\alpha \in \mathbb{R}_+$, $p \in (0,1)$	$\frac{(\alpha)_y}{y!}p^y(1-p)^\alpha$

TABLEAU 1.1 – Distributions de type série entière additive pour $c \in \{-1, 0, 1\}$

Finalement, les distributions inverse Pólya sont données par l'hypergéométrique négative ($\mathcal{NH}_n$), la binomiale négative et la Waring généralisée ($\mathcal{GW}$) pour $c = -1$, $c = 0$ et $c = 1$ respectivement (voir Tableau 1.2). D'une manière similaire, Peyhardi [2023] démontre que ces lois sont stables par composition.

Théorème 1.23 (Peyhardi [2023])**.** Soit $\mathbf{Y} \sim \mathcal{P}_{\Delta_n}^{[c]}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$, alors les marginales Y_j sont stables si $\mathcal{L}$ est donnée par

- $\mathcal{NH}_{|\boldsymbol{\theta}|}(\alpha, \gamma)$ pour $c = -1$, $\gamma \in \mathbb{N}$, $\boldsymbol{\theta} \in \mathbb{N}_+^J$ et $\alpha \in \{0, \ldots, |\boldsymbol{\theta}|\}$;
- $\mathcal{NB}(\alpha, p)$ pour $c = 0$ et tout $\alpha \in \mathbb{R}_+$ et $p \in (0, 1)$;
- $\mathcal{GW}(r, |\boldsymbol{\theta}|, \gamma)$ pour $c = 1$, $\boldsymbol{\theta} \in \mathbb{R}_+^J$ et $r, \gamma \in \mathbb{R}_+$.

Distributions	Paramètres	Fonction de masse
$\mathcal{NH}_n(\alpha, \theta)$	n, $\theta \in \mathbb{N}$ et $\alpha \in \{0, \ldots, \theta\}$	$\frac{\binom{y+\alpha-1}{y}\binom{n+\theta-\alpha-y}{n-y}}{\binom{n+\theta}{n}}$
$\mathcal{GW}(r, \alpha, \beta)$	$r, \alpha, \beta \in \mathbb{R}_+$	$\frac{\mathrm{B}(\alpha+y, \beta+r)}{\mathrm{B}(\alpha, \beta)} \frac{(r)_y}{y!}$

TABLEAU 1.2 – Distributions inverse Pólya pour $c \in \{-1, 1\}$

Relation avec les copules

Comme introduit au début de cette section, il est possible d'utiliser les copules lorsque les marginales sont discrètes. Cette approche possède cependant son lot de défis, dont la perte d'unicité. Néanmoins, il s'avère que le modèle Splitting possède une connexion particulière aux copules. Tel que suggéré par Jones et Marchand [2019], supposons que $\mathcal{S}_{\Delta_n}(\boldsymbol{\theta})$ est un mélange multinomiale, c'est-à-dire $\mathcal{S}_{\Delta_n}(\boldsymbol{\theta}) = \mathcal{M}_{\Delta_n}(\boldsymbol{\pi}) \underset{\boldsymbol{\pi}}{\wedge} \mathcal{S}_{\Delta}(\boldsymbol{\theta})$ où $\mathcal{S}_{\Delta}(\boldsymbol{\theta})$ est une loi sur le simplexe réel positif $\Delta := \{\mathbf{y} \in \mathbb{R}_+^J : |\mathbf{y}| = 1\}$, et $\mathcal{L}(\boldsymbol{\psi})$ est un mélange Poisson, c'est-à-dire $\mathcal{L}(\boldsymbol{\psi}) = \mathcal{P}(\lambda) \underset{\lambda}{\wedge} \mathcal{M}(\boldsymbol{\psi})$ où $\mathcal{M}(\boldsymbol{\psi})$ est une loi sur $\mathbb{R}_+$ (Voir Section 1.3). Supposons également que les variables aléatoires λ et $\boldsymbol{\pi}$ sont indépendantes. Par définition du modèle Splitting et le théorème de Fubini, la distribution de $\mathbf{Y}$ est équivalente à

$$\begin{aligned}\mathbf{Y} \sim \mathcal{S}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi}) &= \left[\mathcal{M}_{\Delta_n}(\boldsymbol{\pi}) \underset{\boldsymbol{\pi}}{\wedge} \mathcal{S}_{\Delta}(\boldsymbol{\theta})\right] \underset{n}{\wedge} \left[\mathcal{P}(\lambda) \underset{\lambda}{\wedge} \mathcal{M}(\boldsymbol{\psi})\right] \\ &= \left[\prod_{j=1}^{J} \mathcal{P}(\lambda\pi_j)\right] \underset{\lambda\boldsymbol{\pi}}{\wedge} \mathcal{M}(\boldsymbol{\psi}) \cdot \mathcal{S}_{\Delta}(\boldsymbol{\theta}),\end{aligned}$$

où on a utilisé le Théorème 1.22 pour la deuxième équivalence. Définissons le vecteur aléatoire $\mathbf{R} = (R_1, \ldots, R_J)$ tel que $\mathbf{R} := \lambda\boldsymbol{\pi}$. On obtient alors que $\mathbf{Y}$ est un mélange

Poisson multivarié, c'est-à-dire

$$\mathbf{Y} \sim \left[\prod_{j=1}^{J} \mathcal{P}(R_j)\right] \underset{\mathbf{R}}{\wedge} \mathcal{R}(\boldsymbol{\psi}, \boldsymbol{\theta})$$

où $\mathcal{R}(\boldsymbol{\psi}, \boldsymbol{\theta})$ est la distribution du produit des variables aléatoires λ et $\boldsymbol{\pi}$. Si $\mathcal{S}_\Delta$ et $\mathcal{M}$ sont absolument continues avec densités $g_{\boldsymbol{\pi}}$ et h_λ respectivement, alors la densité de $\mathcal{R}$ est donnée par

$$f_{\mathbf{R}}(\mathbf{r}) = \frac{h_\lambda(|\mathbf{r}|)}{|\mathbf{r}|^{J-1}} g_{\boldsymbol{\pi}}\left(\frac{\mathbf{r}}{|\mathbf{r}|}\right).$$

La distribution de $\mathbf{R}$ possède plusieurs propriétés intéressantes selon la loi de λ et $\boldsymbol{\pi}$ [c.f. Fang *et al.*, 2013]. En particulier, le vecteur $\mathbf{R}$ est intrinsèquement connecté aux copules.

En effet, le théorème de Sklar [1959] peut être également formulé pour les fonctions de survie jointes. Précisément, pour une fonction de survie jointe $\overline{F}(\mathbf{y}) \coloneqq \mathbb{P}(Y_1 > y_1, \ldots, Y_J > y_J)$ et marginales $\overline{F}_1, \ldots, \overline{F}_J$, il existe une copule de survie $\overline{\mathcal{C}}$ unique sur $\mathrm{Im}(\overline{F}_1) \times \cdots \times \mathrm{Im}(\overline{F}_J)$ telle que $\overline{F}(\mathbf{y}) = \overline{\mathcal{C}}(\overline{F}_1(y_1), \ldots, \overline{F}_J(y_J))$. Dans ce cas, la copule de survie $\overline{\mathcal{C}}$ est définie de manière similaire à la copule. Ainsi, la distribution du vecteur aléatoire $\mathbf{R}$ est caractérisée uniquement par une copule de survie puisqu'on suppose que λ et $\boldsymbol{\pi}$ sont des variables aléatoires continues. En particulier, si $\mathcal{S}_\Delta(\boldsymbol{\theta}) = \mathcal{D}_\Delta(\boldsymbol{\theta})$, la loi de Dirichlet, et $\boldsymbol{\theta} = \mathbf{1}$, McNeil et Nešlehová [2009] démontrent que la copule de survie de $\mathbf{R}$ doit nécessairement être une copule archimédienne. Précisément, si $\mathbb{P}(\lambda \leq 0) = 0$, il existe un générateur ψ tel que

$$\overline{\mathcal{C}}(u_1, \ldots, u_J) = \psi\big(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_J)\big).$$

En fait, la variable aléatoire $\mathbf{R}$ dans ce cas caractérise les copules archimédiennes grâce aux fonctions J-monotones et la transformation de Williamson [c.f. Williamson, 1956; McNeil et Nešlehová, 2009]. Pour le cas général où $\mathcal{S}_\Delta(\boldsymbol{\theta}) = \mathcal{D}_\Delta(\boldsymbol{\theta})$ et $\boldsymbol{\theta} \in \mathbb{R}_+^J$, Gupta et Richards [1987] définissent la distribution de $\mathbf{R}$ comme étant la loi *Liouville multivariée*. Similairement, McNeil et Nešlehová [2010] proposent la *copule Liouville* comme étant la

copule de survie de $\mathbf{R}$. En incorporant ces résultats dans le modèle Splitting, on constate que le cas particulier où

$$\mathbf{Y} \sim \mathcal{DM}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \left[\mathcal{P}(\lambda) \underset{\lambda}{\wedge} \mathcal{M}(\boldsymbol{\psi}) \right]$$

est équivalent au mélange Poisson multivarié

$$\mathbf{Y} \sim \left[\prod_{j=1}^{J} \mathcal{P}(R_j) \right] \underset{\mathbf{R}}{\wedge} \mathcal{R}(\boldsymbol{\psi}, \boldsymbol{\theta}),$$

où les composantes du vecteur $\mathbf{R}$ interagissent selon une copule de survie Liouville.

Moments factoriels

Les moments factoriels sont des valeurs utiles pour représenter les distributions discrètes. Pour une variable aléatoire univariée, le moment factoriel de $Y \in \mathbb{N}$ est l'espérance de la factorielle décroissante d'ordre r, c'est-à-dire

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

Comme présenté dans Johnson *et al.* [2005], la fonction de masse de Y peut être exprimée par

$$\mathbb{P}(Y = n) = \frac{1}{n!} \sum_{k=n}^{\infty} \frac{(-1)^{k-n}}{(k-n)!} \mathrm{E}[(Y)_{(k,-1)}],$$

pour autant que la série converge. Prenons comme exemple $Y \sim \mathcal{P}(\lambda)$, on peut facilement démontrer que

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

Si λ est une variable aléatoire, c'est-à-dire Y est un mélange Poisson, on obtient que le r-ième moment factoriel de Y est le r-ième moment de λ. Les moments factoriels pour les variables multivariées discrètes sont définis de manière similaire. Soit $\mathbf{r} = (r_1, \ldots, r_J) \in \mathbb{N}^J$, le moment factoriel de $\mathbf{Y} \in \mathbb{N}^J$ est l'espérance

$$\mathrm{E}\left[(\mathbf{Y})_{(\mathbf{r},-1)}\right] := \mathrm{E}\left[\prod_{j=1}^{J} (Y_j)_{(r_j,-1)}\right] = (-1)^{|\mathbf{r}|} \mathrm{E}\left[\prod_{j=1}^{J} (-Y_j)_{r_j}\right] =: (-1)^{|\mathbf{r}|} \mathrm{E}\left[(\mathbf{Y})_{\mathbf{r}}\right].$$

Par la même approche, on peut représenter la fonction de masse multivariée du vecteur $\mathbf{Y}$. Il est possible d'obtenir des expressions de ces moments lorsque $\mathbf{Y}$ est Pólya Splitting. En effet, comme on présente au Chapitre 3, on a le résultat suivant.

Théorème 1.24. Soit $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ et $\mathbf{r} \in \mathbb{N}^J$, alors

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

où μ_r est le r-ième moment factoriel de $\mathcal{L}(\boldsymbol{\psi})$.

Covariance et corrélation

On conclut cette section en présentant la covariance et corrélation du modèle Splitting. Soit $\mathbf{Y} \sim \mathcal{P}^{[c]}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi})$ et toute paire Y_i, Y_j, on a à l'aide du Théorème 1.24 que

$$\begin{aligned} \mathrm{Cov}(Y_i, Y_j) &= \mathrm{E}[Y_i Y_j] - \mathrm{E}[Y_i]\mathrm{E}[Y_j] \\ &= \frac{\theta_i \theta_j}{|\boldsymbol{\theta}|^2 (|\boldsymbol{\theta}| + c)} \left[(\mu_2 - \mu_1^2) |\boldsymbol{\theta}| - c\mu_1^2 \right]. \end{aligned} \tag{1.16}$$

Une conséquence directe de (1.16) est que le signe de la covariance est entièrement déterminé par le signe de $(\mu_2 - \mu_1^2)|\boldsymbol{\theta}| - c\mu_1^2$. Puisque cette quantité est indépendante des indices i et j, le signe doit être le même pour toute paire (Y_i, Y_j). De plus, on a par définition des moments factoriels que $\mu_2 - \mu_1^2 = \mathrm{Var}(|\mathbf{Y}|) - \mathrm{E}(|\mathbf{Y}|)$. La dispersion de $\mathcal{L}$ joue donc un rôle important dans la structure de covariance. Prenons comme exemple $\mathbf{Y} \sim \mathcal{DM}_{\Delta_n}(\boldsymbol{\theta}) \underset{n}{\wedge} \mathcal{NB}(\alpha, p)$. Puisque les moments factoriels de la binomiale négative sont donnés par $\mu_r = (\alpha)_r p^r/(1-p)^r$, la covariance (1.16) est donnée, dans ce cas, par

$$\mathrm{Cov}(Y_i, Y_j) = \alpha \frac{\theta_i \theta_j}{|\boldsymbol{\theta}|^2 (|\boldsymbol{\theta}| + 1)} \left(\frac{p}{1-p} \right)^2 (|\boldsymbol{\theta}| - \alpha).$$

Les signes de la covariance sont donc négatifs, nuls ou positifs si α est plus petit, égal ou plus grand que $|\boldsymbol{\theta}|$ respectivement. Notez que le cas où $\alpha = |\boldsymbol{\theta}|$ correspond aux indépendances établies dans le Théorème 1.22.

Finalement, pour la corrélation, on remarque par le Théorème 1.24 que les moments factoriels de $\mathcal{L}$ peuvent être exprimés par les moments factoriels des marginales, c'est-à-dire

$$\mu_r = \frac{(|\boldsymbol{\theta}|)_{(r,c)}}{(\theta_j)_{(r,c)}} \mathrm{E}[(Y_j)_{(r,c)}].$$

Grâce à cette identité, on démontre au Chapitre 3 que la corrélation s'exprime comme

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

où

$$M_i = \frac{\mu_1 \left(1 + \frac{c}{|\boldsymbol{\theta}|}\mu_1\right)}{\mu_2 \left(\frac{\theta_i + c}{|\boldsymbol{\theta}| + c}\right) + \mu_1 \left(1 - \mu_1 \frac{\theta_i}{|\boldsymbol{\theta}|}\right)}.$$

CHAPITRE 2

Asymptotic tail properties of Poisson mixture distributions

L'article présenté dans ce chapitre est un travail conjoint avec mes superviseurs Gwladys Toulemonde, Jean Peyhardi, Éric Marchand et Frédéric Mortier et a été publié dans *Stat*. La référence de l'article est la suivante [Valiquette *et al.*, 2023] :

Samuel Valiquette, Gwladys Toulemonde, Jean Peyhardi, Éric Marchand, & Frédéric Mortier. *Asymptotic tail properties of Poisson mixture distributions. Stat*, 12(1), e622, 2023.

L'article publié peut être consulté avec le lien suivant : `https://doi.org/10.1002/sta4.622`.

Résumé

Les données de comptage sont omniprésentes dans de nombreux domaines d'application et sont généralement surdispersées. Dans ce contexte, les mélanges Poisson forment une solution élégante à ce problème. Cet article s'intéresse à la relation entre le comportement en queue de la distribution de mélange et celui du mélange Poisson résultant. Nous définissons cinq familles de lois de mélange et identifions pour chacune si le mélange Poisson est élément, proche ou éloigné d'un domaine d'attraction maximal. Nous caractérisons également la façon dont le mélange Poisson se comporte de manière similaire à une loi Poisson lorsque la distribution de mélange possède un support fini. Finalement, nous étudions, de manière analytique et numérique, comment la qualité de l'ajustement du mélange Poisson peut être évaluée en examinant le comportement en queue de cette dernière.

Abstract

Count data are omnipresent in many applied fields, often with overdispersion. With mixtures of Poisson distributions representing an elegant and appealing modelling strategy, we focus here on how the tail behaviour of the mixing distribution is related to the tail of the resulting Poisson mixture. We define five sets of mixing distributions and we identify for each case whenever the Poisson mixture is in, close to or far from a domain of attraction of maxima. We also characterize how the Poisson mixture behaves similarly to a standard Poisson distribution when the mixing distribution has a finite support. Finally, we study, both analytically and numerically, how goodness-of-fit can be assessed with the inspection of tail behaviour.

2.1 Introduction

Count data are classically observed in many applied fields such as in actuarial science when evaluating risk and the pricing of insurance contracts [e.g., Bartoszewicz, 2005], in genetics to model the number of genes involved in phenotype variability [e.g., Anders et Huber, 2010] or in ecology to model species abundance [e.g., Wenger et Freeman, 2008]. While Poisson models and regression are well established choices for these type of data, they are not suitable for overdispersed data. To overcome such limitations the use of Poisson mixture models has been proposed. This assumes the Poisson's intensity is no longer an unknown fixed value, but a positive random variable. A variety of mixing distributions has been already proposed [Karlis et Xekalaki, 2005] and classical examples include the gamma distribution [Greenwood et Yule, 1920], the lognormal [Bulmer, 1974] and the Bernoulli [Lambert, 1992]. As demonstrated by Feller [1943], Poisson mixtures are uniquely identifiable by the mixing distribution on the Poisson parameter λ. Therefore, it suffices to take into account the behaviour of the mixing distribution when it comes to adjusting count data with a Poisson mixture model. In particular, the mixing distribution should reflect the tail behaviour of the count data.

The field of extreme value theory allows to analyze such a behaviour through the distribution of maxima. Precisely, the tail behaviour of a random variable can be characterizes by three domains of attraction [Resnick, 1987] : Weibull, Gumbel and Fréchet. Most familiar continuous distributions can be associated to one of these domain of attraction. For discrete distributions, Anderson [1970] identified three different cases. A sample drawn from a discrete distribution is : (i) in a domain of attraction, (ii) "close" to the Gumbel domain of attraction, or (iii) drastically fails to belong to one such that their maxima oscillates between two increasing integers as the sample size grows to infinity. Perline [1998] provided conditions on the mixing distribution such that the Poisson mixture re-

mains in the Fréchet or Gumbel domain of attraction, i.e. case (i). However, they did not report on types of distributions on λ causing the Poisson mixture to satisfy case (ii) or (iii). This article aims to complement their work by identifying conditions on the mixing distribution which allow the Poisson mixture to be associated to the two latter cases. Moreover, we demonstrate that Perline [1998] condition for the Fréchet domain of attraction is not necessary in order for the Poisson mixture to remain in this domain.

This paper is organized as follows. Section 2.2 presents the extreme value theory in the Poisson mixture context and different families of mixing distributions. Using this set of distributions, we identify when the Poisson mixture is in, close to, or far from a domain of attraction. Moreover, we demonstrate that Poisson mixtures satisfying the latter case behave similarly to a standard Poisson distribution. In Section 2.3, we inspect how those three situations can affect the goodness-of-fit when it comes to adjusting count data with a Poisson mixture. Moreover, we explore how one can identify which type of mixing distribution can be adequate by using the generalized Pareto distribution on the excesses. We also study how the closeness to the Gumbel domain of attraction has an impact on identifying such a mixing distribution. Finally, we provide an example where the maxima of a Poisson mixture alternates between two values.

2.2 Poisson mixture tail behaviour

In this section, we present notations and the family of mixing distributions that is studied in this paper. Moreover, preliminary results in extreme value theory are presented and we describe maximum domain of attraction restrictions for discrete distributions. Following this, we elaborate on mixing distributions that allow the Poisson mixture to be in or near a domain of attraction, or to drastically fail to belong in one. Finally, for a Poisson mixture with a finite mixing distribution, we will prove that the asymptotic behaviour

of its probability mass function behaves similarly to that of a Poisson distribution.

2.2.1 Theoretical foundations

In the following, for a Poisson mixture random variable X, i.e. $X \mid \lambda \sim \text{Poisson}(\lambda)$ with λ random, F, $\overline{F}$ and f will denote respectively the cumulative distribution function (cdf), the survival function, and the probability density function (pdf) for the mixing λ. Similarly, F_M, $\overline{F}_M$ and P_M will denote respectively the cdf, the survival function, and the probability mass function (pmf) of the resulting Poisson mixture X. Moreover, in this paper, we restrict the mixing distributions on λ to those with a support equal to $(0, x_0)$ for $x_0 \in \mathbb{R}_+ \cup \{\infty\}$. Finally, we require the notion of a slowly varying function $C(x)$ on $\mathbb{R}_+$, defined by the property : for every $t \in \mathbb{R}_+$, $C(tx) \sim C(x)$, where $g(x) \sim h(x)$ means that $\lim_{x\to\infty} \frac{g(x)}{h(x)} = 1$ for functions g and h.

The tail behaviour of the Poisson mixture can be studied using extreme value theory. Such a statistical approach analyzes how the maxima of F_M stabilizes asymptotically. For a general distribution G, the theory says that G belongs to a domain of attraction if there exist two normalizing sequences $a_n > 0$ and b_n such that $G^n(a_n x + b_n)$ converges to a non-degenerate distribution when n tends to infinity [Resnick, 1987]. Such a non-degenerate distribution can only be the generalized extreme value distribution given by

$$\lim_{n\to\infty} G^n(a_n x + b_n) = \begin{cases} \exp\left[-(1+\gamma x)^{-1/\gamma}\right] & \text{for } 1+\gamma x > 0 \text{ with } \gamma \neq 0\,; \\ \exp\left[-e^{-x}\right] & \text{for } x \in \mathbb{R} \text{ with } \gamma = 0. \end{cases} \tag{2.1}$$

The three possible domains of attraction are named Weibull, Gumbel and Fréchet for $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$ respectively, and will be denoted by $\mathcal{D}_-$, $\mathcal{D}_0$ and $\mathcal{D}_+$. Accordingly, we will write $G \in \mathcal{D}$ where $\mathcal{D}$ is one of the three domains. Necessary and sufficient conditions for G to be in a domain of attraction have been established by Gnedenko [1943]. While most common continuous distributions can be associated to a domain of

attraction, this is not always the case for discrete random variables. Indeed, a necessary condition for a discrete distribution G to be in a domain of attraction is the long-tailed property [Anderson, 1970] defined by

$$\overline{G}(n+1) \sim \overline{G}(n). \tag{2.2}$$

Well known discrete distributions, such as Poisson, geometric and negative binomial, do not satisfy the above property. However, Anderson [1970] and Shimura [2012] showed that if a discrete distribution verifies

$$\overline{G}(n+1) \sim L\overline{G}(n), \tag{2.3}$$

for $L \in (0,1)$, then G is, in a sense, "close" to the Gumbel domain. More precisely, Shimura [2012] showed that property (2.3) implies that G is the discretization of a unique continuous distribution belonging to $\mathcal{D}_0$. On the other hand, Anderson [1970] showed that there exist a sequence b_n and $\alpha > 0$ such that

$$\begin{aligned} \limsup_{n\to\infty} G^n(x+b_n) &\leq \exp\left(-e^{-\alpha x}\right) \\ \liminf_{n\to\infty} G^n(x+b_n) &\geq \exp\left(-e^{-\alpha(x-1)}\right) \end{aligned}$$

if and only if $\overline{G}(n+1) \sim e^{-\alpha}\overline{G}(n)$. Therefore, the supremum and infimum limits of $G^n(x+b_n)$ are bounded by two Gumbel distributions under condition (2.3). The geometric and negative binomial distributions are two such examples. Finally, if the discrete distribution is a Poisson, or more generally such that

$$\lim_{n\to\infty} \frac{\overline{G}(n+1)}{\overline{G}(n)} = 0, \tag{2.4}$$

then no sequence b_n can be found such that the the supremum and infimum limits of $G^n(x+b_n)$ are bounded by two different Gumbel distributions. For this case, Anderson [1970] showed that for $Y_i \overset{iid}{\sim} G$, there exists a sequence of integers I_n such that

$$\lim_{n\to\infty} P\left(\max_{1\leq i\leq n} Y_i = I_n \text{ or } I_n + 1\right) = 1 \tag{2.5}$$

if and only if (2.4) is satisfied. Therefore, the maximum of such discrete distribution oscillates between two integers asymptotically.

2.2.2 Poisson mixtures categories

Since Poisson mixture distributions are discrete, they are constrained to the long-tailed property (2.2) in order to have a domain of attraction. Otherwise, they may be close to the Gumbel domain or with a maximum alternating between two integers. Since a Poisson mixture is uniquely identifiable by the distribution on λ [Feller, 1943], it follows that its tail behaviour depends on the latter. Therefore, we seek to identify what conditions on the distribution of λ allow the Poisson mixture distributions to satisfy one of the three equations (2.2), (2.3) or (2.4). In the following, we will establish that Poisson mixtures with F in $\mathcal{D}_+$ or $\mathcal{D}_-$ will satisfy equations (2.2) and (2.4) respectively, but for mixing distributions in $\mathcal{D}_0$, the Poisson mixture may satisfy one of the three limits depending on their behaviour. We require the following definitions and notations.

Definition 2.1. A distribution F has an **exponential tail** if for all $k \in \mathbb{R}$, there is a $\beta > 0$ such that for $x \to \infty$

$$\overline{F}(x+k) \sim e^{-\beta k}\overline{F}(x). \tag{2.6}$$

Definition 2.2. A distribution F satisfies the **Gumbel hazard condition** if its density f has a negative derivative for all x in some left neighborhood of $\{+\infty\}$, $\lim_{x\to\infty} \frac{d}{dx}\left[\frac{1-F(x)}{f(x)}\right] = 0$ (the 3rd Von Mises Condition) and $\lim_{x\to\infty} \frac{x^\delta f(x)}{1-F(x)} = 0$ for some $\delta \geq \frac{1}{2}$.

Using Definitions 2.1 and 2.2, we focus on three distinct subsets of $\mathcal{D}_0$. Firstly, distributions satisfying one of these definitions are in the Gumbel domain of attraction, see Shimura [2012] and Resnick [1987]. Secondly, some distributions with finite tail are in $\mathcal{D}_0$ (e.g. Gnedenko, 1943). Based on these three cases, let $\mathcal{D}_0^{\mathcal{E}}$, $\mathcal{D}_0^{\mathcal{H}}$ and $\mathcal{D}_0^{\mathcal{F}}$ denote respectively

the classes of $F \in \mathcal{D}_0$ satisfying Definition 2.1, Definition 2.2, and with finite tail. These subsets of $\mathcal{D}_0$ are disjoint by the following Proposition.

Proposition 2.1. The sets $\mathcal{D}_0^{\mathcal{E}}$, $\mathcal{D}_0^{\mathcal{H}}$ and $\mathcal{D}_0^{\mathcal{F}}$ are disjoint.

Proof. Since $\mathcal{D}_0^{\mathcal{E}}$ and $\mathcal{D}_0^{\mathcal{H}}$ represent distributions with an infinite tail, they are both disjoint from $\mathcal{D}_0^{\mathcal{F}}$. To establish that $\mathcal{D}_0^{\mathcal{E}}$ and $\mathcal{D}_0^{\mathcal{H}}$ are disjoint, we first note that for any distribution $F \in \mathcal{D}_0^{\mathcal{H}}$, the density must exist and it must be ultimately decreasing by Definition 2.2. Therefore, if $F \in \mathcal{D}_0^{\mathcal{E}}$ and these conditions are not satisfied, then $F \notin \mathcal{D}_0^{\mathcal{H}}$. Finally, suppose $F \in \mathcal{D}_0^{\mathcal{E}}$, its density exists and is ultimately monotone. It is sufficient to show that the condition on the hazard rate function in Definition 2.2 is not satisfied for such a distribution. As noticed in Cline [1986], F has an exponential tail if and only if $\overline{F}(\ln x) = C(x)x^{-\beta}$ for some $\beta > 0$ and slowly varying function C. By the monotone density theorem presented in Theorem 1.7.2. in Bingham *et al.* [1987], we then have $f(x) \sim C(e^x)\beta e^{-\beta x}$. Using these properties, one has for all $\delta > 0$ that

$$\lim_{x\to\infty} \frac{x^\delta f(x)}{\overline{F}(x)} = \lim_{x\to\infty} \frac{x^\delta f(x)}{C(e^x)e^{-\beta x}} = \beta \lim_{x\to\infty} x^\delta = \infty,$$

showing that the Gumbel hazard condition (Definition 2.2) is not satisfied.

□

Although these subsets are disjoint, they do not form a partition of $\mathcal{D}_0$. Indeed, the Weibull distribution with cdf $F(x) = 1 - e^{-\left(\frac{x}{\beta}\right)^\alpha}$ is neither in $\mathcal{D}_0^{\mathcal{H}}$ or $\mathcal{D}_0^{\mathcal{E}}$ when $\alpha \notin (0, 1/2) \cup \{1\}$. This distribution belongs to a broader subset of $\mathcal{D}_0$ named Weibull tail which intersects with $\mathcal{D}_0^{\mathcal{H}}$ and $\mathcal{D}_0^{\mathcal{E}}$; see Gardes et Girard [2013] for more details. We now discriminate between properties (2.2), (2.3) or (2.4) with respect to the domain of attraction of λ.

Theorem 2.1. Let F_M be a Poisson mixture with λ distributed according to a cdf F and supported on $(0, x_0)$ with $x_0 \in \mathbb{R}_+ \cup \{\infty\}$. Then for any integer $k \geq 1$, there is a $\beta > 0$

such that

$$\lim_{n\to\infty} \frac{\overline{F}_M(n+k)}{\overline{F}_M(n)} = \begin{cases} 1 & \text{if } F \in \mathcal{D}_+ \cup \mathcal{D}_0^{\mathcal{H}}, \\ (1+\beta)^{-k} & \text{if } F \in \mathcal{D}_0^{\mathcal{E}}, \\ 0 & \text{if } F \in \mathcal{D}_- \cup \mathcal{D}_0^{\mathcal{F}}. \end{cases}$$

Proof. (A) $\lim_{n\to\infty} \frac{\overline{F}_M(n+k)}{\overline{F}_M(n)} = 1$: The result for $\mathcal{D}_0^{\mathcal{H}}$ is directly established by Perline [1998]. For $F \in \mathcal{D}_+$, a necessary and sufficient condition is that $\overline{F}(x) = C(x)x^{-\alpha}$ with $\alpha > 0$ [Gnedenko, 1943]. In fact, $C(x)$ must be locally bounded since $\overline{F}$ is bounded. As presented in Karlis et Xekalaki [2005], the survival function of the mixture is given by

$$\overline{F}_M(x) = \int_0^\infty \frac{\lambda^{\lfloor x \rfloor} e^{-\lambda}}{\lfloor x \rfloor!}(1 - F(\lambda))d\lambda = \frac{\Gamma(\lfloor x \rfloor - \alpha + 1)}{\Gamma(\lfloor x \rfloor + 1)} \int_0^\infty \underbrace{\frac{\lambda^{\lfloor x \rfloor - \alpha} e^{-\lambda}}{\Gamma(\lfloor x \rfloor - \alpha + 1)}}_{g(x,\lambda)} C(\lambda)d\lambda$$

for x such that $\lfloor x \rfloor - \alpha > 0$. By the definition of the Gamma function, $\int_0^\infty g(x,\lambda)d\lambda = 1$, and then for $0 \le a < b \le \infty$, $\phi \in (-1, 1)$, and Stirling's formula, we have

$$\int_a^b \lambda^\phi g(x,\lambda)d\lambda \le \int_0^\infty \lambda^\phi g(x,\lambda)d\lambda = \frac{\Gamma(\lfloor x \rfloor - \alpha + \phi + 1)}{\Gamma(\lfloor x \rfloor - \alpha + 1)} \sim \lfloor x \rfloor^\phi.$$

By Theorem 4.1.4 in Bingham *et al.* [1987], we can conclude that $\overline{F}_M$ is such that

$$\overline{F}_M(x) \sim C(\lfloor x \rfloor)\frac{\Gamma(\lfloor x \rfloor - \alpha + 1)}{\Gamma(\lfloor x \rfloor + 1)} \sim C(\lfloor x \rfloor)\lfloor x \rfloor^{-\alpha}.$$

Furthermore, since $\lfloor x \rfloor \sim x$, $C(\lfloor x \rfloor) \sim C(x)$ using the Karamata representation of C [Resnick, 1987]. Therefore $F_M \in \mathcal{D}_+$ and $\overline{F}_M(n+k) \sim \overline{F}_M(n)$.

(B) $\lim_{n\to\infty} \frac{\overline{F}_M(n+k)}{\overline{F}_M(n)} = (1+\beta)^{-k}$: Since F has an exponential tail, then $\overline{F}(x) = C(e^x)e^{-\beta x}$ for some $\beta > 0$. Using a similar argument as in Theorem 4.1.4 in Bingham *et al.* [1987], we can prove that

$$\overline{F}_M(n) \sim \frac{C(e^n)}{(1+\beta)^{n+1}}.$$

Therefore,

$$\lim_{n\to\infty} \frac{1 - F_M(n+k)}{1 - F_M(n)} = (1+\beta)^{-k} \lim_{n\to\infty} \frac{C(e^{n+k})}{C(e^n)} = (1+\beta)^{-k}.$$

(C) $\lim_{n\to\infty} \frac{\overline{F}_M(n+k)}{\overline{F}_M(n)} = 0$: Because $\overline{F}_M(n) = \int_0^{x_0} \frac{\lambda^n e^{-\lambda}}{n!}(1 - F(\lambda))d\lambda$, the result as above follows since

$$\begin{aligned}\frac{\overline{F}_M(n+k)}{\overline{F}_M(n)} &= \frac{1}{\prod_{i=1}^{k}(n+i)} \frac{\int_0^{x_0} \lambda^{n+k} e^{-\lambda}(1 - F(\lambda))d\lambda}{\int_0^{x_0} \lambda^n e^{-\lambda}(1 - F(\lambda))d\lambda} \\ &\leq \frac{x_0^k}{\prod_{i=1}^{k}(n+i)} \to 0 \text{ when } n \to \infty.\end{aligned}$$

□

Theorem 2.1 establishes that if $F \in \mathcal{D}_+$, then $F_M \in \mathcal{D}_+$ which improves the result of Perline [1998]. Indeed, Perline's proof requires the 1st Von Mises condition [Resnick, 1987] to prove a similar result. By relaxing such a condition, we proved that any mixing distributions in $\mathcal{D}_+$ allows the Poisson mixture to remain in this domain of attraction. Analogous to this property, Shimura [2012] showed that any discretization of a continuous distribution in $\mathcal{D}_+$ preserves the domain of attraction. Considering the Poisson mixture as a discretization operator, we obtain another example where the Fréchet domain of attraction is preserved. A broad set of mixing distributions in $\mathcal{D}_+$ can be found, for example the Fréchet, folded-Cauchy, Beta type II, inverse-Gamma, or the Gamma/Beta type II mixture [Irwin, 1968]. Unfortunately, examples are scarce for distributions in $\mathcal{D}_0^{\mathcal{H}}$. Indeed the asymptotic behaviour of the hazard rate function in Definition 2.2 is quite restrictive. Examples include the lognormal, the Benktander type I and II [Kleiber et Kotz, 2003], and the Weibull distributions, with further restrictions on the parameters for the latter two cases. These type of distributions do not encompass cases like the Gamma, even though the associated mixing distribution belongs to $\mathcal{D}_0$, because it does not satisfy the additional condition on the hazard rate function. The class $\mathcal{D}_0^{\mathcal{E}}$ allows to describe such a mixing distribution. It includes a broad class of elements among others Gamma, Gamma/Gompertz, exponential, exponential logarithmic, inverse-Gaussian and

the generalized inverse-Gaussian. As previously mentioned these distributions are in the Gumbel domain of attraction but, from Theorem 2.1, the resulting Poisson mixtures do not belong to any domain of attraction. However, we can quantify how close such Poisson mixtures are to the Gumbel domain of attraction. Indeed, if $\beta \to 0$ then $\frac{1-F_M(n+1)}{1-F_M(n)} \to 1$, i.e. it approaches a long-tailed distribution. Finally, when F has a finite tail, i.e. $F \in \mathcal{D}_- \cup \mathcal{D}_0^{\mathcal{F}}$, the Poisson mixture cannot be close to any domain of attraction by Theorem 2.1.

2.2.3 Asymptotic behaviour for $F \in \mathcal{D}_-$

To shed light on why the last limit in Theorem 2.1 is null, we complete this section by studying the asymptotic behaviour of the pmf P_M when F is in $\mathcal{D}_-$. Willmot [1990] studied such a behaviour when the Poisson mixture has a mixing distribution with a particular exponential tail. This result is presented in the following Proposition.

Proposition 2.2 (Willmot [1990])**.** Let F_M be a Poisson mixture with λ distributed according to a distribution F such that its density is

$$f(x) \sim C(x)x^{\alpha}e^{-\beta x},$$

where C is a locally bounded and slowly varying function on $\mathbb{R}_+$, and for some $\alpha \in \mathbb{R}$ and $\beta > 0$. Then the pmf P_M is such that

$$P_M(n) \sim C(n)n^{\alpha}(1+\beta)^{-(n+\alpha+1)}.$$

Proposition 2.2 indicates that when the density f behaves similarly to a Gamma distribution, then the pmf P_M behaves like a negative binomial pmf multiplied by a regular varying function. As previously mentioned, the negative binomial is an example of a distribution where equation (2.3) is satisfied. This provides additional clarification on why the limit associated with an exponential tail in Theorem 2.1 converges to a value between 0 and 1. In the following Theorem, a similar conclusion is presented when $F \in \mathcal{D}_-$.

Theorem 2.2. Let F_M be a Poisson mixture with λ distributed according to a distribution $F \in \mathcal{D}_-$. Then there exists an $\alpha > 0$ such that

$$\overline{F}_M(n) \sim \Gamma(\alpha+1)C(n)n^{-\alpha}\left(\frac{x_0^{n+1}}{(n+1)!}e^{-x_0}\right).$$

Proof. Using the integral representation of $\overline{F}_M$, we have

$$\overline{F}_M(n) = \int_0^{x_0} \frac{\lambda^n e^{-\lambda}}{n!}(1-F(\lambda))d\lambda = \frac{x_0^{n+1}}{n!}\int_0^{\infty} \frac{\lambda^n}{(\lambda+1)^{n+2}} e^{-\frac{x_0\lambda}{\lambda+1}}\left(1-F\left(\frac{x_0\lambda}{\lambda+1}\right)\right)d\lambda$$

where the transformation $\lambda \mapsto \frac{\lambda}{x_0-\lambda}$ has been applied. By adapting the necessary and sufficient condition for the Weibull domain of attraction [Gnedenko, 1943], which is $F \in \mathcal{D}_-$ if and only if $x_0 < \infty$ and $1 - F\left(\frac{x_0 x}{x+1}\right) = C(x)x^{-\alpha}$ for C a locally bounded function and slowly varying and $\alpha > 0$, we obtain

$$\overline{F}_M(n) = \frac{x_0^{n+1}}{n!}\int_0^{\infty} \frac{\lambda^{n-\alpha}}{(\lambda+1)^{n+2}} C(\lambda) e^{-\frac{x_0\lambda}{\lambda+1}} d\lambda$$

then using the fact that the Beta function is such that

$$\mathrm{B}(a,b) = \int_0^{\infty} \frac{t^{a-1}}{(t+1)^{a+b}}dt,$$

a similar argument as in Theorem 2.1 provides that

$$\begin{aligned}
\overline{F}_M(n) &\sim \frac{x_0^{n+1}}{n!}\mathrm{B}(n-\alpha+1,\alpha+1)C(n)e^{-\frac{x_0 n}{n+1}} \\
&\sim \frac{x_0^{n+1}e^{-x_0}}{n!}C(n)\frac{\Gamma(n-\alpha+1)\Gamma(\alpha+1)}{\Gamma(n+2)} \\
&\sim \Gamma(\alpha+1)C(n)n^{-\alpha}\left(\frac{x_0^{n+1}e^{-x_0}}{(n+1)!}\right).
\end{aligned}$$

□

With the asymptotic behaviour in Theorem 2.2, a similar result can be established for P_M.

Corollary 2.1. Let F_M be a Poisson mixture with λ distributed according to a cdf $F \in \mathcal{D}_-$. Then the pmf P_M is such that

$$P_M(n) \sim \Gamma(\alpha+1)C(n)n^{-\alpha}\left(\frac{x_0^n}{n!}e^{-x_0}\right).$$

Proof. Since $P_M(n) = \overline{F}_M(n-1) - \overline{F}_M(n)$, then

$$\lim_{n\to\infty} \frac{P_M(n)}{\Gamma(\alpha+1)C(n)n^{-\alpha}\left(\frac{x_0^n}{n!}e^{-x_0}\right)} = \lim_{n\to\infty} \frac{C(n-1)(n-1)^{-\alpha}}{C(n)n^{-\alpha}} - \lim_{n\to\infty} \frac{x_0}{n+1} = 1.$$

□

This result provides a fresh perspective on why the limit in Theorem 2.1 converges to 0 for a mixing distribution with a finite support. Indeed, as previously mentioned, the Poisson distribution is an example such that the limit (2.4) is satisfied. From Theorem 2.2 and Corollary 2.1, $\overline{F}_M$ and P_M behave like a Poisson distribution with mean x_0 multiplied by a regular varying function. Intuitively, the mixing distribution does not put weight everywhere on $\mathbb{R}_+$, so the tail of F_M cannot satisfy equation (2.2).

2.3 Numerical Study

This section illustrates the practical implications of the theoretical results previously obtained. In particular, we highlight how the mixing distribution impacts the adjustment, how the statistical evaluation of tail distributions of count data may help to select a mixing distribution, and how the maxima of Poisson mixtures with finite mixing distribution behave asymptotically.

2.3.1 Impact of mixing distribution choice on goodness of fit

To illustrate how the tail behaviour of λ affects the model adjustment, we simulated 100 samples of different Poisson mixtures with size $n = 250$ using the (i) Fréchet(α, β), (ii) lognormal(μ, σ), (iii) Gamma(α, β), and (iv) Uniform$(0, x_0)$ distributions on λ with densities

(i) $f(x) = \frac{\alpha}{x}\left(\frac{x}{\beta}\right)^{-\alpha} e^{-\left(\frac{x}{\beta}\right)^{-\alpha}}$, $\alpha > 0$, $\beta > 0$;

(ii) $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$ $\mu \in \mathbb{R}$, $\sigma > 0$;

(iii) $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, $\alpha > 0$, $\beta > 0$;

(iv) $f(x) = \frac{\mathbb{1}_{(0,x_0)}(x)}{x_0}$,

each one being a representative of four out of five type of mixing distributions we encountered. Respectively, they are representative of elements in $\mathcal{D}_+$, $\mathcal{D}_0^{\mathcal{H}}$, $\mathcal{D}_0^{\mathcal{E}}$, and in $\mathcal{D}_-$. Moreover, the parameter γ from equation (2.1) associated to (i) and (iv) are respectively $\gamma = 1/\alpha$, $\gamma = -1$ and $\gamma = 0$ for (ii) and (iii). For each sample, the Poisson mixture is fitted with the same four distributions and the best model is kept using a Bayesian framework. This is done using the language `R` [R Core Team, 2021] and the `rstan` [Stan Development Team, 2020] package to estimate the hyperparameters by MCMC. The best model is then kept using the highest *posterior* model probability. Those probabilities are approximated using the bridge sampling computational technique [Meng et Wong, 1996] and the dedicated `R` package `Bridgesampling` [Gronau *et al.*, 2020]. All results are based on the following priors : a Gamma$(1, 1)$ distribution for positive parameters and a Normal$(0, 1)$ for real parameters. Moreover, we simulated for each sample four MCMCs with 10,000 iterations each in order to ensure reasonable convergence for parameter estimation and for the *posterior* model probabilities. Results are presented in Table 2.1.

Mixing class	Mixing distribution	Fréchet	Lognormal	Gamma	Uniform
$\mathcal{D}_+$	Fréchet(1,1)	**89**	11	0	0
	Fréchet(2,1)	**80**	18	2	0
$\mathcal{D}_0^{\mathcal{H}}$	Lognormal(1,1)	5	**89**	6	0
	Lognormal(0,1)	9	**69**	23	0
$\mathcal{D}_0^{\mathcal{E}}$	Gamma(2,1)	1	22	**73**	4
	Gamma(2,2)	1	23	**54**	22
$\mathcal{D}_-$	Uniform(0,10)	0	0	26	**74**
	Uniform(0,5)	0	1	38	**61**

TABLEAU 2.1 – Selected model frequencies for each Poisson mixture simulation, with the highest frequency in bold.

The Poisson-Fréchet mixtures stand out the most since their tail is heavier than any other of the distributions. The only competing model seems to be the Poisson-lognormal which has a heavier tail than an exponential type distribution, but lighter than the Fréchet. The variance also influences what model is selected. Indeed, for example, the lognormal(0,1) has a lower variance compared to the lognormal(1,1). In the former mixture, the Gamma seems to be able to compete against the lognormal, which is not the case for the latter. Interestingly, the Fréchet mixing distribution is selected sparingly for lognormal data even when the variance gets larger. This fact remains true for the rest of Table 2.1 since the Fréchet distribution has a much heavier tail. By Theorem 2.1, we know that the Gamma distribution can get close to the Gumbel domain of attraction. From Table 2.1, we see that the lognormal is a significant competitor for both simulations, which reflects the closeness to $\mathcal{D}_0$. However, when the rate parameter is equal to 2, the mean and variance decrease and the uniform becomes another chosen option. This can be explained by the fact that $\frac{\overline{F}_M(n+1)}{\overline{F}_M(n)}$ is closer to 0 when n grows to infinity. Finally, since the uniform has a finite tail, only the Gamma can compete and, again, larger the variance the less the Gamma is selected. Based on each case, we see a diagonal effect from the heavier tail to the finite tail.

2.3.2 Identifying the domain of attraction

In order to identify what domain of attraction a random variable belongs to, one can uses the peaks-over-threshold (POT) method [Coles, 2001]. This technique involves the distribution of the excesses defined by $Y - u|Y > u$, for a suitable choice of u. Pickands [1975], Balkema et Haan [1974] showed that Y belongs to a domain of attraction if and only if the distribution of the excesses converges weakly to a generalized Pareto distribution (GPD) as u tends to the right endpoint of the distribution of Y. In such cases, the corresponding cdf is given by

$$H_{\gamma,\sigma}(y) = \begin{cases} 1 - \left(1 + \gamma \frac{y}{\sigma}\right)^{-1/\gamma} & \text{if } \gamma \neq 0, \\ 1 - \exp\left(-\frac{y}{\sigma}\right) & \gamma = 0, \end{cases} \tag{2.7}$$

with support $\mathbb{R}_+$ if $\gamma \geq 0$ or $\left[0; -\frac{\sigma}{\gamma}\right]$ if $\gamma < 0$, where $\gamma \in \mathbb{R}$ and $\sigma > 0$ are respectively shape and scale parameters. Moreover, the γ parameter is the same as in equation (2.1). Therefore, fitting a GPD to the excesses of a sample can inform us on the domain of attraction the underlying distribution belongs to. Better yet, excesses of count data can inform us whether or not a Poisson mixture distribution belongs to a known domain of attraction and, if so, which one. Therefore, analyzing the discrete excesses can indicate what type of mixing distribution generates the Poisson mixture. Indeed, by Theorem 2.1, if the discrete excesses belong to a domain of attraction, then a mixing distribution F should be in $\mathcal{D}_+ \cup \mathcal{D}_0^{\mathcal{H}}$. Otherwise, F should either have an exponential or finite tail.

From a practical point of view, the study of discrete excesses may justify a choice of model. For example, one may hesitate between adjusting a Poisson-lognormal or a negative binomial for their count data. In order to study how useful the discrete excesses can be, various Poisson mixtures have been simulated. Here, we fixed the sample size to $n = 1000$, the threshold u to be the 95th or 97.5th empirical quantiles, and simulated 1000 samples for each mixing distribution. For each sample, the discrete excesses are extracted, and the `evd` R package [Stephenson, 2002] is used to estimate the GPD parameters by

maximum likelihood. Based on these estimations, the modified Anderson Darling test for the goodness-of-fit is applied. Finally, for the samples such that the GPD appears to be adequate, we test $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$. To do so, we fit these two models, evaluate the corresponding log likelihoods $\mathcal{L}_1$ and $\mathcal{L}_0$, and conclude with the deviance statistic $D = 2(\mathcal{L}_1 - \mathcal{L}_0)$ which follows approximately a χ^2_1 distribution under suitable conditions [Coles, 2001]. Results are presented in Table 2.2.

Mixing Class	Mixing distribution	u	Average number of access	GPD Rejection	Test $\gamma = 0$ not rejected
$\mathcal{D}_+$	Fréchet(1,1)	95	48.727	0.069	0.014
		97.5	24.685	0.051	0.158
	Fréchet(2,1)	95	41.915	0.777	0.170
		97.5	21.746	0.177	0.615
$\mathcal{D}_0^{\mathcal{H}}$	Lognormal(1,1)	95	46.750	0.126	0.720
		97.5	23.644	0.037	0.845
	Lognormal(0,1)	95	41.914	0.697	0.257
		97.5	21.685	0.142	0.790
$\mathcal{D}_0^{\mathcal{E}}$	Gamma(2,1)	95	36.200	0.704	0.045
		97.5	18.876	0.245	0.502
	Gamma(2,2)	95	38.015	0.833	0.052
		97.5	16.988	0.392	0.311
$\mathcal{D}_-$	Uniform(0,10)	95	39.124	0.641	0.028
		97.5	18.999	0.296	0.390
	Uniform(0,5)	95	35.161	0.679	0.059
		97.5	18.087	0.369	0.255

TABLEAU 2.2 – Average number of excesses, rejection rate for the GPD, and non-rejection rate of $H_0 : \gamma = 0$ with $n = 1000$ and $u =$ 95th or 97.5th empirical quantile.

Firstly, we notice that even if the Fréchet and lognormal distributions are in $\mathcal{D}_+$ and $\mathcal{D}_0^{\mathcal{H}}$ respectively, the Fréchet(2,1) and lognormal(0,1) cases lead to a high rejection rate for the 95th quantile threshold. However, when both cases are simulated with a threshold u equal to the 97.5th quantile, the rate of GPD rejection diminishes. Therefore, it seems that the threshold choice has a great impact. Moreover, when u is the 97.5th quantile,

the estimation of γ is not significantly different to 0 for 79 % of the samples of the lognormal(0,1). However, 61.5 % of the samples of the Fréchet are also significantly null. Secondly, as noted by Hitz *et al.* [2017], the discrete excesses need a certain amount of variability in order to have a smooth adjustment to the GPD. Since the lognormal(1,1) has a greater variance and the Fréchet(1,1) doesn't have a finite expectation, this explains why these cases are well adjusted to the GPD. Finally, both Gamma and uniform cases have GPD rejection rates as expected. Interestingly, the uniform distribution is rejected at a lesser rate then the Gamma. Again, this can be explained by the greater variance for the uniform than the Gamma simulations.

Also, the Gamma(2,1) leads to a lower rate of rejection than the Gamma(2,2), which is reasonable since the former is closer to $\mathcal{D}_0$ than the latter by Theorem 2.1. Indeed, if the limit in Theorem 2.1 $(1+\beta)^{-1}$ approaches 0, the GPD rejection rate for the Poisson mixtures should increase. Inversely, the rejection rate should decrease when $(1+\beta)^{-1}$ approaches 1. To further analyze this, we simulated Poisson mixtures with a Gamma(2, β) mixing density and let the parameter β vary from 0.1 to 8, the quantity $(1+\beta)^{-1}$ thus varying between 1/9 and 10/11. For each value of β, we simulated 500 samples of size $n = 1000$ from the Poisson mixture, fix the threshold u to the 95th empirical quantile, and calculate the proportion of samples where the GPD is rejected with type I error $\alpha = 0.05$. Results are presented in Figure 2.1. We can see that indeed the proportion decreases when $(1+\beta)^{-1}$ moves towards 1. Between 0 and 0.5, the rejection proportion oscillates between 0.5 and 1. This can be explained by the fact that the number of discrete excesses also oscillates when β increases, which affects the power of the test.

To adjust for the problems related to the discreteness of the excesses, it would be interesting to transform them into continuous variables. As demonstrated by Shimura [2012], a Poisson mixture with $F \in \mathcal{D}_0^{\mathcal{H}}$ is a random variable that originates from an unique continuous distribution in $\mathcal{D}_0$ that has been discretized. If one can identify such a continuous

distribution associated to the discrete excesses when the GPD is rejected, then it would be reasonable to use an exponential tail mixing distribution. A jittering technique consiting of adding random noise to data has been proposed for different discrete contexts [Nagler, 2018; Coeurjolly et Rousseau Trépanier, 2020]. A plausible approach would be a jittering for the GPD test in order to adequately identify the type of mixing distribution associated to the discrete excesses.

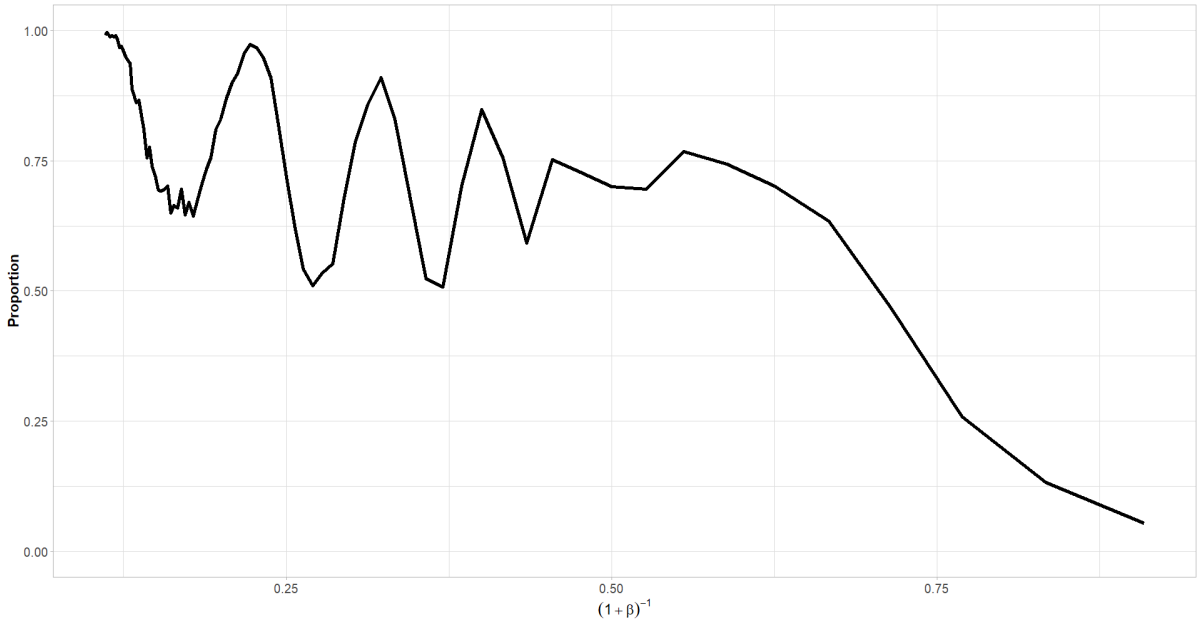


FIGURE 2.1 – Proportion of Gamma(2, β) Poisson mixture samples (size $n = 1000$) where the GPD has been rejected ($\alpha = 0.05$) for the excesses (u = 95th empirical quantile) as a function of $(1+\beta)^{-1}$.

2.3.3 Maxima for Poisson mixtures with finite tail mixing distribution

By Theorem 2.1, if F has bounded support $(0, x_0)$, then the Poisson mixture is short tailed, i.e. $\frac{\overline{F_M}(n+1)}{\overline{F_M}(n)} \to 0$ as $n \to \infty$. Therefore, according to Anderson [1970], there exists

a sequence of integers I_n such that equation (2.5) is satisfied. Moreover, by Corollary 2.1, the pmf P_M asymptotically behaves like a Poisson distribution and, as mentioned, the Poisson is the primary example where its maximum oscillates between two integers. Kimber [1983] and Briggs *et al.* [2009] study how the sequence I_n can be approximated for the Poisson distribution and showed that it grows slowly when $n \to \infty$. Since P_M behaves like the Poisson when F is in $\mathcal{D}_-$, the sequence I_n should also grow slowly. To visualise this behaviour, we simulated Poisson mixtures with $\lambda \sim x_0\text{Beta}(\alpha, \beta)$. We fixed $\alpha = 2$, $x_0 = 5$, and for $n \in \{10, 10^2, 10^3, 10^4\}$, we simulated 10000 samples of F_M with size n and recorded the maximum for each sample. With these maxima, we calculated the empirical probabilities, and repeated for $\beta \in \{1/4, 1/2, 1, 2\}$. Figure 2.2 reports on the empirical and theoretical pmf of the simulations and the maxima of n Poisson variables with mean x_0 respectively. Interestingly, the greater β becomes, the slower the sequence I_n increases.

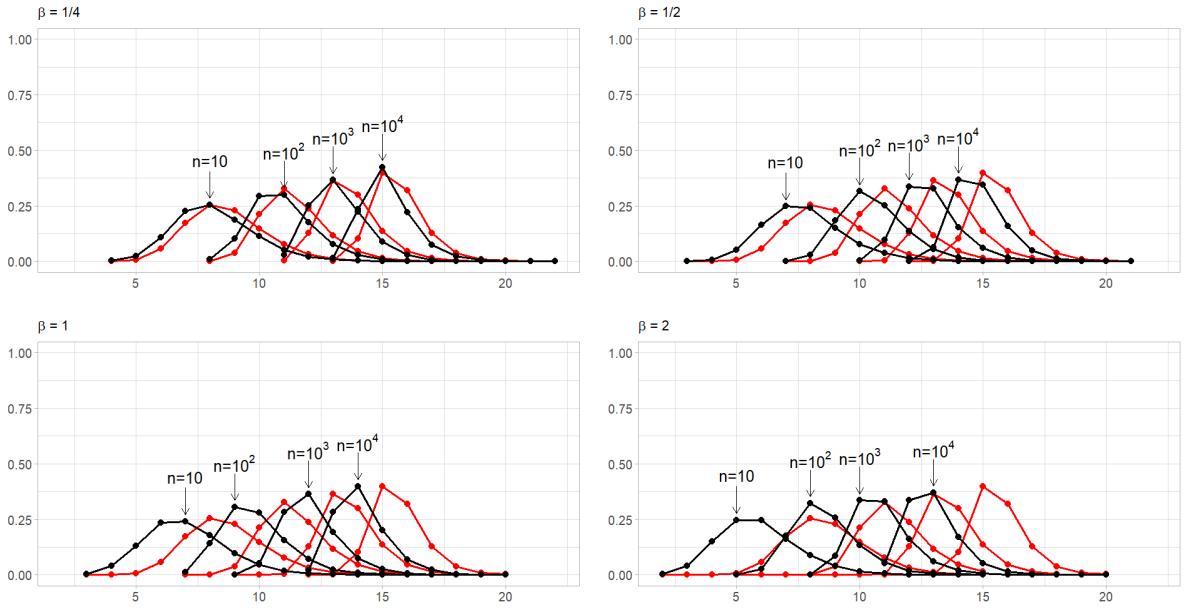


FIGURE 2.2 – Maximum distributions of Poisson mixture with $\lambda \sim x_0\text{Beta}(2, \beta)$ (black) and Poisson(x_0) (red) with $x_0 = 5$, $\beta \in \{1/4, 1/2, 1, 2\}$ and $n \in \{10, 10^2, 10^3, 10^4\}$.

Indeed, when $\beta = 1/4$, the probability distribution of the maxima looks similar to that

of a Poisson(x_0). For $\beta = 2$, the distribution for the Poisson mixture drastically shifts to the left. This can be explained using Corollary 2.1. Indeed, we can show that P_M here is such that

$$P_M(n) \sim \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} n^{-\beta} \left(\frac{x_0^n e^{-x_0}}{n!} \right),$$

and when β approaches 0, then only the pmf of the Poisson(x_0) remains. From another point of view, the density of the $x_0\text{Beta}(\alpha, \beta)$ approaches a Dirac on x_0, so the Poisson mixture approaches a simple Poisson distribution.

2.4 Conclusion and outlook

Overdispersed count data are commonly observed in many applied fields and Poisson mixtures are appealing to model such data. However, the choice of the appropriate mixing distribution is a difficult task relying mainly on empirical approaches related to modelers subjectivity or on intensive computational techniques combined with goodness-of-fit test or information criteria. In this paper, we showed that such a choice should respect the relation between the tail behaviour of λ and the discrete data. Indeed, if a distribution F is in the Fréchet domain of attraction or satisfies the Gumbel hazard condition given by Definition 2.2, then the discrete data should be in the same domain of attraction. Otherwise, an exponential or a finite tail should be chosen. Moreover, Theorem 2.1 established that Poisson mixtures with $F \in \mathcal{D}_0$ need to be separated into three subsets : $\mathcal{D}_0^{\mathcal{E}}$, $\mathcal{D}_0^{\mathcal{H}}$ and $\mathcal{D}_0^{\mathcal{F}}$. Both subsets $\mathcal{D}_0^{\mathcal{E}}$ and $\mathcal{D}_0^{\mathcal{H}}$ have distributions belonging to a larger subset named Weibull tail [Gardes et Girard, 2013]. It would be interesting to generalize Theorem 2.1 with this familly of mixing distributions.

To identify whether the data distribution comes from a domain of attraction or not, we have studied the discrete excesses and their adjustment by the GPD. Some difficulties occurred due to the discrete nature of the data. Solutions that could be explored are the

use of techniques like the jittering or the use of discrete analogues of the GPD like the discrete generalized Pareto or the generalized Zipf distribution presented in Hitz *et al.* [2017]. These approaches should help identify whether λ has a exponential tail or not. However, one could consider testing whether or not the distribution on λ has a bounded support. To elaborate such a test, an interesting avenue would be to use Theorem 2.2 and Corollary 2.1, which state that the Poisson mixture with a finite mixing distribution should behave similarly to a Poisson with mean x_0.

In the field of extreme value theory, our Theorem 2.2 and the result of Willmot [1990] in Proposition 2.2 may provide an approach to finding normalizing sequences such that the Poisson mixture belongs to a domain of attraction. Indeed, Anderson *et al.* [1997] showed that if the Poisson's mean λ depends on the sample size and increases with a certain rate, then it is possible to find normalizing sequences a_n and b_n such that the distribution is in the Gumbel domain of attraction. If λ does not depend on the sample size, then no such sequence can be found. A similar result has been proved by Nadarajah et Mitov [2002] for the negative binomial when α is fixed and β approaches 0. Since Theorem 2.2 and Proposition 2.2 showed that Poisson mixtures with finite or exponential tail mixing distribution resemble the Poisson or the negative binomial respectively, one could exploit these asymptotic properties to generalize the results of Anderson *et al.* [1997] and Nadarajah et Mitov [2002] with various Poisson mixtures like the Poisson-inverse-Gaussian or Poisson-Beta. Similarly, generalizing the results of Kimber [1983] and Briggs *et al.* [2009] concerning the sequence I_n for the maxima of Poisson random variables should also be explored.

CHAPITRE 3

Tree Pólya Splitting distributions for multivariate count data

L'article présenté dans ce chapitre est un travail conjoint avec mes superviseurs Jean Peyhardi, Gwladys Toulemonde, Éric Marchand et Frédéric Mortier. La référence de l'article est la suivante [Valiquette *et al.*, 2024] :

Samuel Valiquette, Jean Peyhardi, Gwladys Toulemonde, Éric Marchand, & Frédéric Mortier. *Tree Pólya Splitting distributions for multivariate count data*. ⟨hal-04563659⟩. 2024.

L'article peut être consulté avec le lien suivant : `https://hal.science/hal-04563659`.

Résumé

Dans cet article, nous développons une nouvelle classe de distributions multivariées adaptées à des données de comptage, dénommée Tree Pólya Splitting. Cette classe résulte de la combinaison d'une distribution univariée et de distributions multivariées singulières le long d'un arbre de partition connu. Comme nous allons le montrer, ces distributions sont flexibles, permettant notamment la modélisation de dépendances complexes (positives, négatives ou nulles) au niveau des variables observées. Plus précisément, nous présentons les propriétés théoriques des distributions Tree Pólya Splitting en nous focalisant principalement sur les lois marginales, les moments factoriels et les structures de dépendance (covariance et corrélations). L'abondance de 17 espèces de trichoptères relevée sur 49 sites est utilisée pour, d'une part, illustrer les propriétés théoriques développées dans cet article sur un cas concret, et d'autre part, montrer l'intérêt de ce type de modèles, notamment en les comparant à des approches classiques en écologie ou en microbiome.

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.

3.1 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 et 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 et Li, 2013; Wang et 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 et 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 et Marchand [2019] proposed the *sums and shares* model where $\mathbf{Y} = (Y_1, \dots, Y_J) \in \mathbb{N}^J$ 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 et 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 et Li, 2013]. The generalized Dirichlet-multinomial, introduced by Connor et Mosimann [1969], is considered by Tang et Chen [2018] with a focus on zero-inflation. Another example is the Dirichlet-tree multinomial model proposed by Dennis [1991] and employed by Wang et 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 et 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 et 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 et 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.

3.2 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 et Marchand [2019], Peyhardi *et al.* [2021], and Peyhardi [2023] for further details.

3.2.1 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 $\Delta_n \coloneqq \{\mathbf{y} \in \mathbb{N}^J : |\mathbf{y}| = n\}$. 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

$$(\theta)_{(n,c)} = \begin{cases} 1 & \text{if } n = 0 \\ \theta(\theta + c) \dots (\theta + (n-1)c) & \text{if } n \geq 1. \end{cases} \tag{3.1}$$

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

$$\mathbf{p}_{|\mathbf{Y}|=n}(\mathbf{y}) = \frac{n!}{(|\boldsymbol{\theta}|)_{(n,c)}} \prod_{j=1}^J \frac{(\theta_j)_{(y_j,c)}}{y_j!}, \tag{3.2}$$

for $c \in \{-1, 0, 1\}$ and parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ [Eggenberger et 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 3.1. (Pólya Splitting distribution) A random vector $\mathbf{Y} = (Y_1, \ldots, Y_J) \in \mathbb{N}^J$ 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

$$\mathbf{p}(\mathbf{y}) = \mathbf{p}(|\mathbf{Y}| = n) \left[\frac{n!}{(|\boldsymbol{\theta}|)_{(n,c)}} \prod_{j=1}^J \frac{(\theta_j)_{(y_j,c)}}{y_j!} \right], \tag{3.3}$$

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 $|\boldsymbol{\theta}| \geq m$.

3.2.2 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

$$Y_j \sim \mathcal{P}^{[c]}_n(\theta_j, |\boldsymbol{\theta}_{-j}|) \underset{n}{\wedge} \mathcal{L}(\boldsymbol{\psi}). \tag{3.4}$$

Notice that the univariate distribution $\mathcal{L}(\boldsymbol{\psi})$ in (3.4) 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

$$\mathbf{p}(y) \propto \frac{\alpha^y}{y!}(\theta)_{(y,c)}, \tag{3.5}$$

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 3.1 for each distribution represented in terms of α, θ and c.

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

TABLEAU 3.1 – 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 $\mathbf{r} = (r_1,\dots,r_J) \in \mathbb{N}^J$, 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 3.1. 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 $\mathcal{L}(\boldsymbol{\psi})$.

3.2.3 Covariance and dispersion

Using Property 3.1, the covariance between Y_i and Y_j for $i \neq j$ in the Pólya Splitting distribution is given by

$$\mathrm{Cov}(Y_i, Y_j) = \frac{\theta_i \theta_j}{|\boldsymbol{\theta}|^2 (|\boldsymbol{\theta}| + c)} \left[(\mu_2 - \mu_1^2) |\boldsymbol{\theta}| - c\mu_1^2 \right]. \tag{3.6}$$

Here, we are particularly interested in the signs of the covariances. Clearly, the sign of (3.6) 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 3.2. 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 3.2 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 $\boldsymbol{\theta}$.

3.2.4 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 3.1 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 3.3. 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 et 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 3.3, we can generalize their result to other values of $\boldsymbol{\theta}$.

Property 3.4. 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 $\theta_j/(\theta_j + 1)$.

3.3 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 et Marchand [2019] results. We also show how the correlations of the Tree Pólya can indeed take various signs.

3.3.1 Definitions and Notations

Let $\mathfrak{T} = (\mathcal{N}, \mathcal{E})$ 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 3.2 (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 3.3 (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 3.1.

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 $\text{Path}_{\{4\}}^{A} = (\{9\}, \{9,10\}, \{8,9,10\}, A)$.

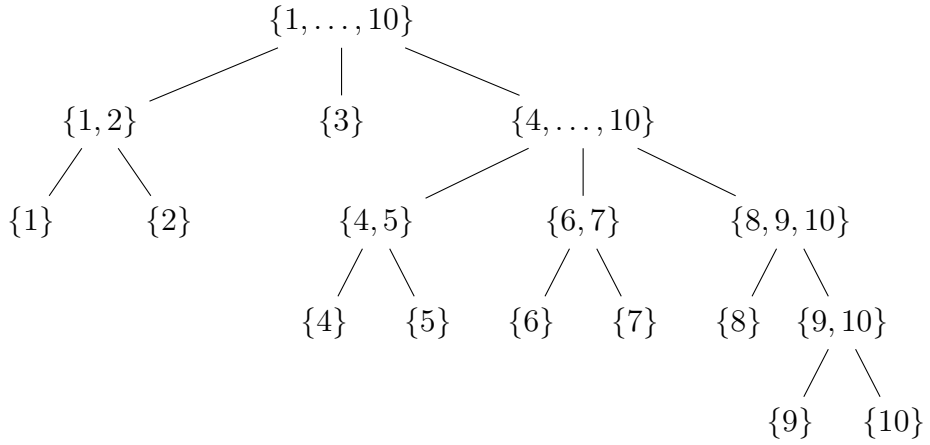


FIGURE 3.1 – Example of a partition tree with 10 leaves.

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 3.4 (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.

$$\mathbf{p}(\mathbf{y}) = \mathbf{p}(|\mathbf{Y}| = n) \prod_{A\in\mathfrak{J}} \frac{n_A!}{(|\boldsymbol{\theta}_A|)_{(n_A,c_A)}} \prod_{C\in\mathfrak{C}_A} \frac{(\theta_C)_{(n_C,c_A)}}{n_C!}, \tag{3.7}$$

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

Notice that if $\mathfrak{J} = \Omega$ in Definition 3.4, then the basic Pólya Splitting p.m.f. (3.3) is recovered in (3.7). 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 3.1, Figure 3.2 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.

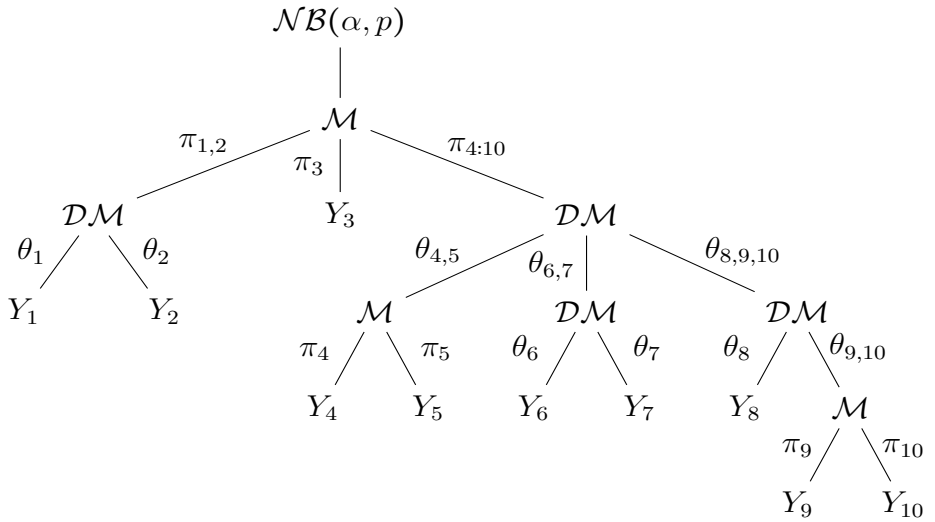


FIGURE 3.2 – Example of Tree Pólya Splitting distribution based on the partition tree in Figure 3.1 and $\mathcal{L} = \mathcal{NB}(\alpha, p)$

.

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 et 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 3.3). 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.

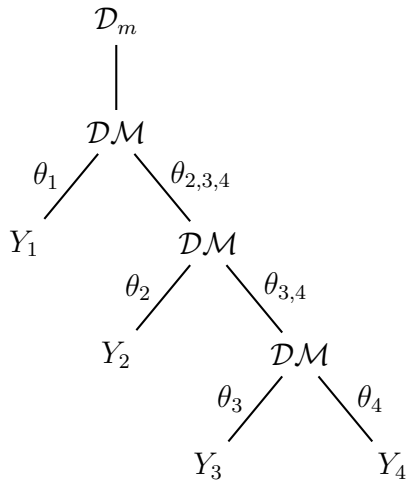


FIGURE 3.3 – Generalized Dirichlet-multinomial distribution represented by a Tree Pólya Splitting model with $\mathcal{L}(\psi) = \mathcal{D}_m$, the Dirac distribution at m.

3.3.2 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 3.2, 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.

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 (3.4). 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 3.1 (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

$$Y_j \sim \bigwedge_{k=1}^{K} \mathcal{P}_{n_k}^{[c_{A_k}]}\left(\theta_{A_{k-1}}, |\boldsymbol{\theta}_{A_k \setminus A_{k-1}}|\right) \underset{n_K}{\wedge} \mathcal{L}(\boldsymbol{\psi}), \tag{3.8}$$

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

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

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 3.1.

Proposition 3.2 (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 (3.8) but with $A_k \in \mathrm{Path}_A$. If $A = \Omega$, then $K = 0$ and $|\mathbf{Y}| \sim \mathcal{L}(\boldsymbol{\psi})$.

Finally, Proposition 3.2 can be used to obtain multivariate marginal distributions that are consistent with the whole tree structure.

Proposition 3.3 (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 3.2, $\tilde{\mathfrak{T}}$ 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 3.2. Using Proposition 3.1, the marginal is given by

$$Y_6 \sim \mathcal{BB}_{n_1}(\theta_6, \theta_7) \underset{n_1}{\wedge} \mathcal{BB}_{n_2}(\theta_{6,7}, \theta_{4,5} + \theta_{8,9,10}) \underset{n_2}{\wedge} \mathcal{B}_{n_3}(\pi_{4:10}) \underset{n_3}{\wedge} \mathcal{NB}(\alpha, p) \tag{3.9}$$

as represented by the path in Figure 3.4.

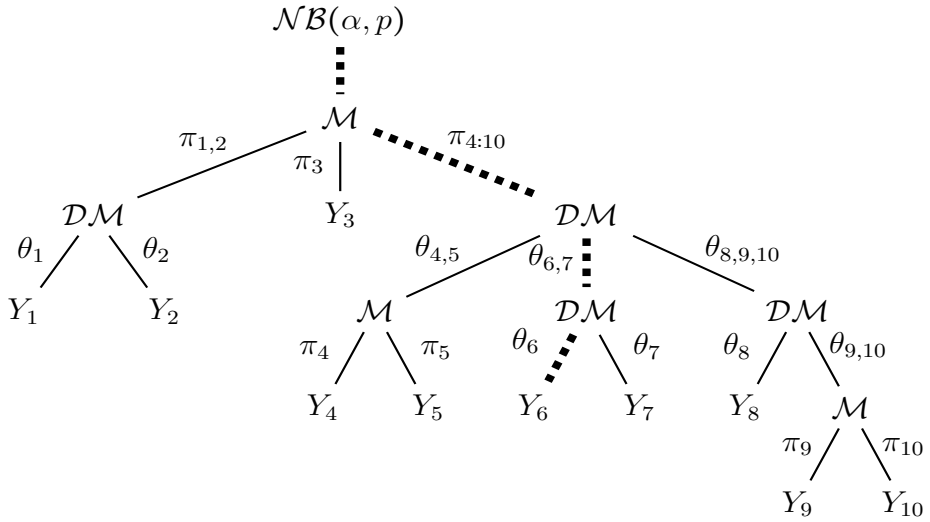


FIGURE 3.4 – Path representation of the marginal Y_6 in our running example.

In fact, equation (3.9) 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 3.4. 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 (3.9), or any marginal in Figure 3.4, it is sufficient to study the p.m.f. of the general composition

$$X \sim \left[\bigwedge_{k=1}^{K} \mathcal{BB}_{n_k}(\alpha_k, \beta_k)\right] \underset{n_K}{\wedge} \mathcal{NB}(\alpha, p). \tag{3.10}$$

For $K = 1$, Jones et Marchand [2019] showed that the p.m.f. of (3.10) 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 et Gupta, 1984], we have the following proposition.

Proposition 3.5. 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 (3.10), 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 3.2, 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 (3.9) can be obtained using Propositions

3.4 and 3.5. 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 3.5 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.

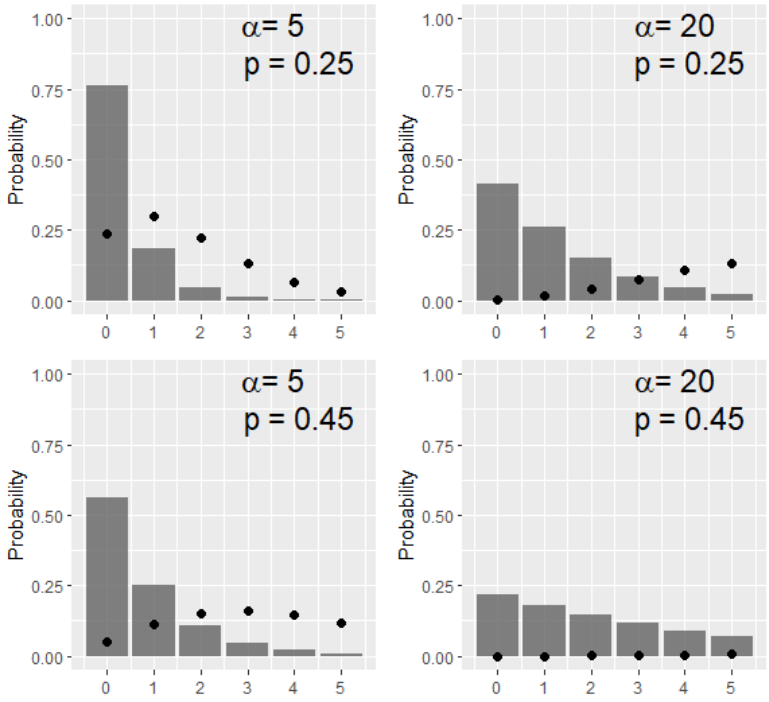


FIGURE 3.5 – P.m.f. of Y_6 (Bar) and $\mathcal{NB}(\alpha, p)$ (Points) with $\theta_6 = 3$, $\theta_7 = \theta_{6,7} = 1$, $\theta_{4,5} + \theta_{8,9,10} = 2$, $\pi_{4:10} = 0.75$, $\alpha \in \{5, 20\}$ and $p = \{0.25, 0.45\}$.

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 3.2. Previously, the binomial splits were absorbed by the negative binomial in accordance to Proposition

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

$$X \sim \bigwedge_{k=1}^{K} \mathcal{P}_{n_k}^{[c_k]}(\theta_k, \tau_k) \underset{n_K}{\wedge} \mathcal{D}_m \tag{3.11}$$

where the parameters are given in Proposition 3.4. Using similar techniques, we prove the following.

Proposition 3.6. Let X be distributed as (3.11) 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 $\gamma = \prod_{k=1}^{K-M} \pi_k$.

Factorial moments

Factorial moments of the Pólya Splitting distribution were determined by Property 3.1 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 3.7. 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 3.7 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 3.1. 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 3.4. Corollary 3.1 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 $\mu_r = (-1)^r(-m)_r$.

Covariance and Correlation

If two leaves are siblings in the partition tree, then their covariance is simply equation (3.6) for a Pólya Splitting model with univariate distribution $\mathcal{L}(\boldsymbol{\psi})$ given by Proposition 3.2. 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 3.7, it can be shown that $\mathrm{Cov}(Y_i, Y_j)$ is proportional to the covariance at their common ancestor.

Proposition 3.8. 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

$$\mathrm{Cov}(Y_i, Y_j) = \frac{\gamma_i \gamma_j}{\gamma_{C_i} \gamma_{C_j}} \mathrm{Cov}(|\mathbf{Y}_{C_i}|, |\mathbf{Y}_{C_j}|), \tag{3.12}$$

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 3.2. 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 3.1, 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 3.9. The covariance (3.12) is given by

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

where $\delta_S = \prod_{k=1}^{K} \frac{\theta_{A_{k-1}} + c_{A_k}}{|\boldsymbol{\theta}_{A_k}| + c_{A_k}}$ with $A_k \in \operatorname{Path}_S$, γ_ℓ defined in Proposition 3.8, and μ_r the r-th factorial moment of $\mathcal{L}$.

Notice that the previous results in Proposition 3.8 and 3.9 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 3.2. 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

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

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 3.1 for the standard deviations and Proposition 3.9 for the covariance, we have the following result.

Proposition 3.10. 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 3.8 and 3.9 respectively.

Running example

Returning to our running example, let us calculate the covariance between Y_6 and Y_9. As presented in Figure 3.6, the ancestor node is given by $S = \{4, \ldots, 10\}$ with $C_6 = \{6, 7\}$ and $C_9 = \{8, 9, 10\}$. By Propositions 3.8 and 3.9, 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

$$\begin{aligned}
\operatorname{Cov}(Y_6, Y_9) &= \left[\frac{\theta_6}{\theta_6 + \theta_7} \right] \left[\pi_9 \frac{\theta_{9,10}}{\theta_8 + \theta_{9,10}} \right] \operatorname{Cov}(Y_4 + Y_5, Y_8 + Y_9 + Y_{10}) \\
&= \alpha \left(\frac{p}{1-p} \right)^2 \left[(\alpha + 1) \left(\frac{|\boldsymbol{\theta}_S|}{|\boldsymbol{\theta}_S| + 1} \right) - \alpha \right] \gamma_6 \gamma_9 \\
&= \alpha \left(\frac{p}{1-p} \right)^2 \left(\frac{|\boldsymbol{\theta}_S| - \alpha}{|\boldsymbol{\theta}_S| + 1} \right) \gamma_6 \gamma_9. \qquad (3.13)
\end{aligned}$$

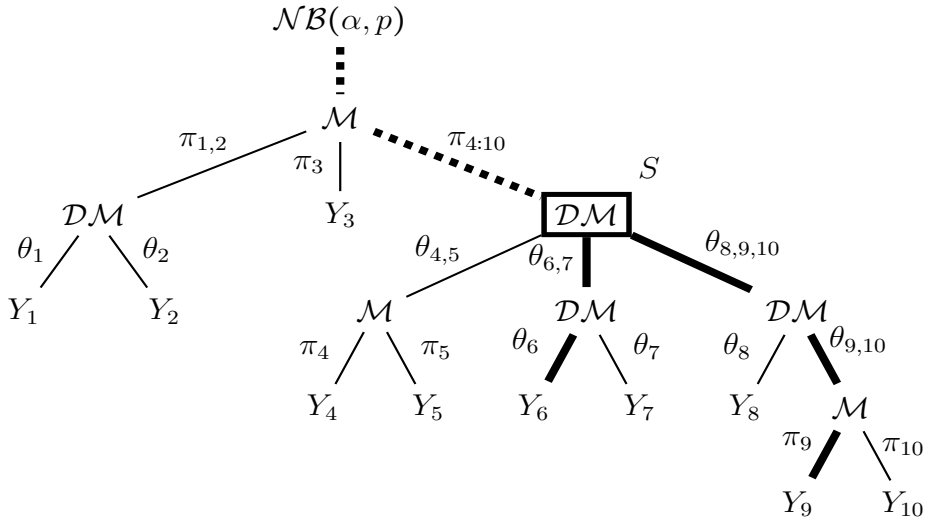


FIGURE 3.6 – Covariance between Y_6 and Y_9 in our running example

Notice that since $\mathcal{L} = \mathcal{NB}(\alpha, p)$, the overdispersion is preserved throughout the tree by Property 3.2. 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 (3.13), 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.

$$\begin{aligned} \alpha &= 10 & \pi_{4:10} &= 0.6 & \theta_6 &= 0.8 \\ p &= 0.95 & \theta_{4,5} &= 3 & \theta_7 = \theta_8 &= 1 \\ \pi_{1,2} = \pi_9 &= 0.3 & \theta_{6,7} = \theta_{8,9.10} &= 3.5 & \theta_{9,10} &= 2.5 \\ \pi_3 &= 0.1 & \pi_4 = \pi_5 &= 0.5 & \pi_{10} &= 0.7 \\ \theta_1 = \theta_2 &= 1.5 \end{aligned} \tag{3.14}$$

Using Proposition 3.10, its correlation is presented in Figure 3.7.

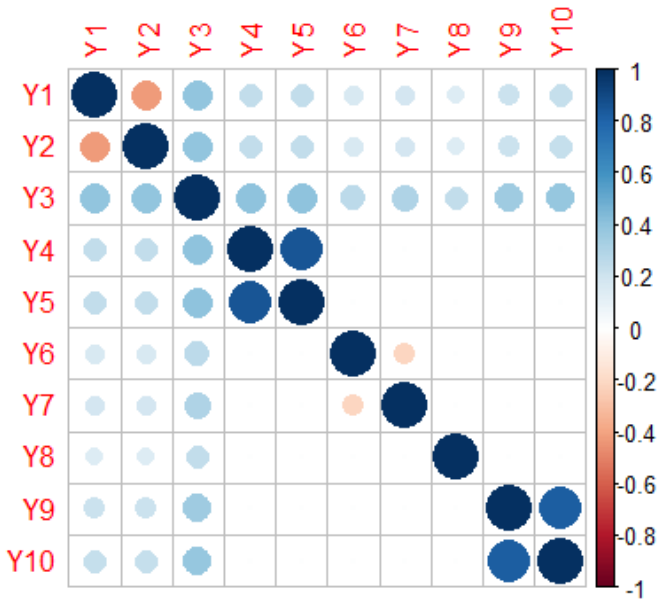


FIGURE 3.7 – Correlation plot of our running example with parameters (3.14) and $\mathcal{L} = \mathcal{NB}(\alpha, p)$.

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 3.2. 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 3.10, the correlation is presented in Figure 3.8 for the Dirac at m.

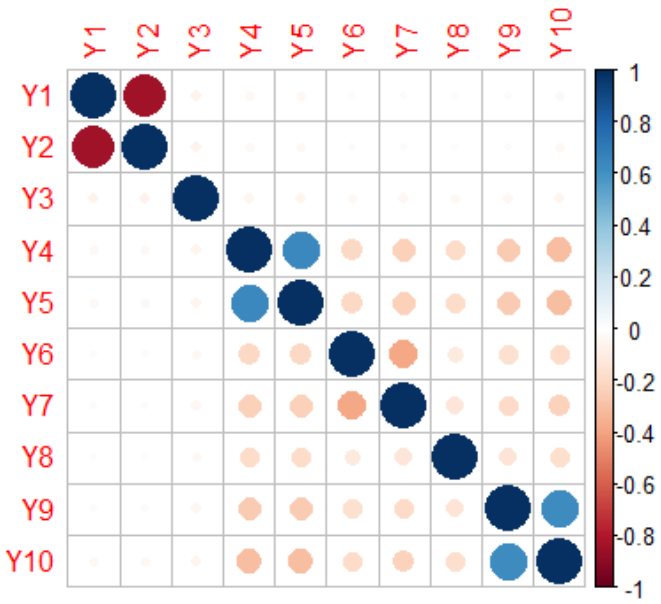


FIGURE 3.8 – Correlation plot of our running example with parameters (3.14) and $\mathcal{L} = \mathcal{D}_{100}$.

Log-likelihood decomposition

A direct consequence of Definition 3.4 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 (3.7) is given by

$$\log\big[\mathbf{p}(|\mathbf{Y}| = n)\big] + \sum_{A \in \mathfrak{I}} \left(\sum_{C \in \mathfrak{C}_A} \log(\theta_C)_{(n_C, c_A)} - \log(|\boldsymbol{\theta}_A|)_{(n_A, c_A)} \right) + \text{constant}. \tag{3.15}$$

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.

3.4 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 et 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 et 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 et 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 et 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 $j \in \{1,\dots,J\}$, then the Dirichlet-multinomial converges to a multinomial with parameters $\boldsymbol{\pi} = (\pi_1,\dots,\pi_J)$.

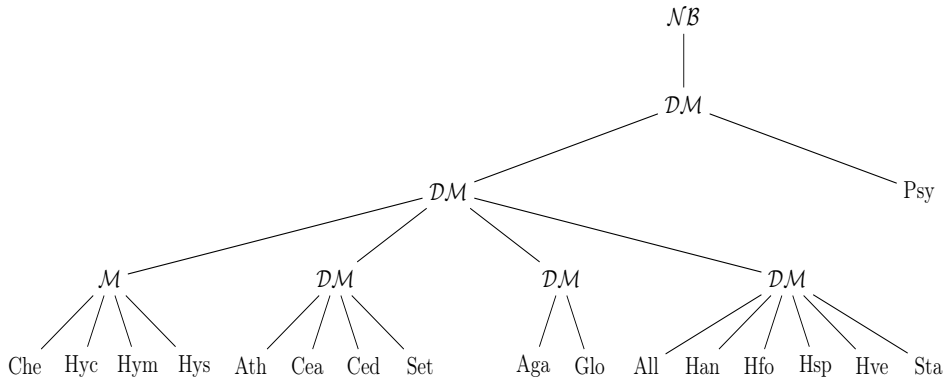


FIGURE 3.9 – Tree Pólya Splitting fitted to the Trichoptera data set with a fixed partition tree.

With these criteria, we have the Tree Pólya model presented in Figure 3.9. 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 3.9, 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 3.10.

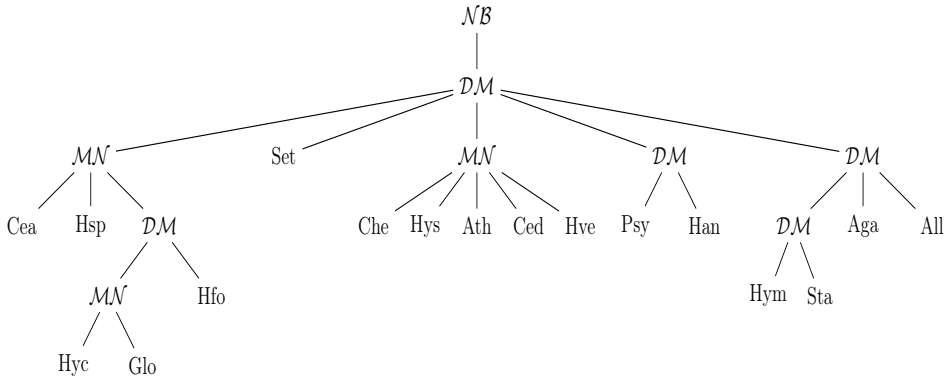


FIGURE 3.10 – Tree Pólya Splitting fitted to the Trichoptera data set with a partition tree search.

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 3.2. Finally, the correlations can be easily calculated by combining Proposition 3.10 and the estimated parameters. See Figure 3.11.

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 3.9)	23	2465.85
Partition tree search (Figure 3.10)	23	2380.77

TABLEAU 3.2 – Fitted models to the Trichoptera data set

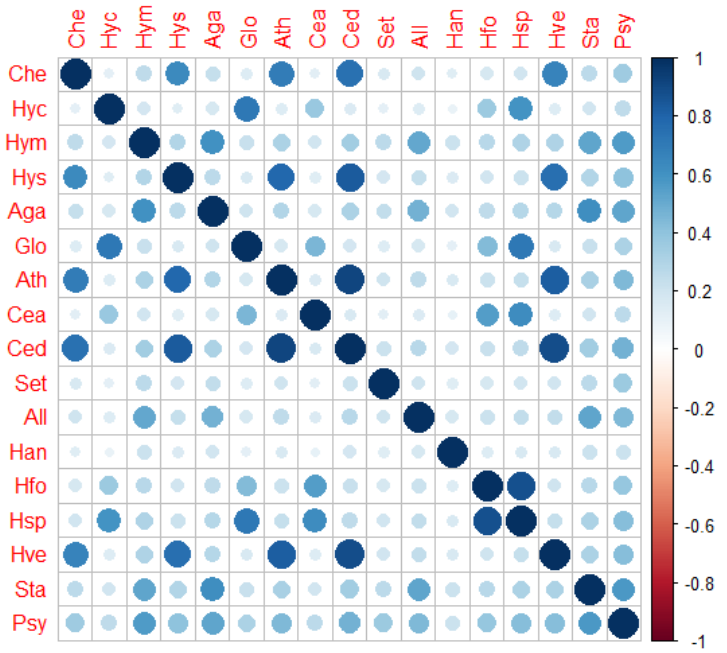


FIGURE 3.11 – Correlation plot of the Tree Pólya Splitting fitted to the Trichoptera data set with a partition tree search.

3.5 Discussion and perspectives

The simplicity and versatility of the Pólya Splitting model proposed by Jones et 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 et 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 probabilis-

tic 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 et 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 et Tian [2015], Santana *et al.* [2022], and Zeng *et al.* [2023], have been proposed. Similarly, Tang et 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.

Appendix - Proofs

Property 3.1

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

Lemma 3.1. 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 3.2

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 3.3

As presented, the covariance is such that

$$\mathrm{Cov}(Y_i, Y_j) = \frac{\theta_i \theta_j}{|\boldsymbol{\theta}|^2}\left[\frac{|\boldsymbol{\theta}|\,\mu_2}{|\boldsymbol{\theta}| + c} - \mu_1^2\right]. \tag{3.16}$$

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 (3.16) 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 3.4

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 3.3, 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 3.1, 3.2 and 3.3

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 (3.4) 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 3.4

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 3.5

By the argument presented in the proof of Proposition 3.4, 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 et 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 3.2. 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

$$\bigwedge_{k=1}^{K} \mathcal{BB}_{n_k}(\alpha_k, \beta_k). \tag{3.17}$$

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 (3.17) is given by

$$G(z) = {}_{K+1}F_K\left[\begin{matrix} -n_K,\ \boldsymbol{\alpha} \\ \boldsymbol{\alpha}+\boldsymbol{\beta} \end{matrix}; 1-z\right]. \tag{3.18}$$

Then for $K+1$ and by conditioning on the last K terms of (3.17), 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 3.2 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 (3.18) 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 (3.18) 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 et 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 3.6

Using a similar argument as in Proposition 3.4, 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 (3.18) and Lemma 3.2, 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 3.7

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 3.1

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

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 3.8

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 3.1 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 3.9

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 3.1 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 (3.6), 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 3.8. □

Proposition 3.10

We only need to calculate the variance of Y_ℓ for $\ell = i, j$. Again, by Corollary 3.1, the result follows since

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

□

CHAPITRE 4

Conclusion et perspectives

Modéliser des données discrètes, et en particulier de comptage, reste un enjeu méthodologique. Par exemple, les changements climatiques ou les pressions humaines influencent de manière évidente la dynamique des écosystèmes. Or comprendre l'influence de ces facteurs repose notamment sur la mise en relation entre les données d'abondance des espèces et leur environnement. Comment prendre en compte l'aspect muti-espèces d'un écosystème ? Comment tenir compte d'événements extrêmes dans la distribution des espèces ? Ce travail de thèse a eu pour objectif de répondre à certains de ces défis. Dans le cadre univarié, nous avons proposé une approche originale pour mieux prendre en compte les valeurs extrêmes. Dans le cadre multivarié, nous proposons une nouvelle classe de modèles basée sur une structure d'arbre de partitionnement qui permet une modélisation flexible des dépendances entre les observations. En particulier, nous avons étendu les résultats théoriques de Willmot [1990] et Perline [1998] pour les valeurs extrêmes dans le cadre des mélanges Poisson et généralisé les modèles Splitting introduits par Peyhardi et Fernique [2017], Jones et Marchand [2019] et Peyhardi *et al.* [2021]. Ces travaux ouvrent diverses perspectives de recherche. Nous proposons de développer celles-ci dans

les quelques pages suivantes. Pour commencer, nous présentons un travail en cours portant sur les représentations graphiques probabilistes de certains Tree Pólya Splitting. Cette perspective est suffisamment développée pour que nous puissions en présenter les détails. Notre objectif, à la suite de cette thèse, est de publier une version sous forme d'article scientifique. Ensuite, nous soulignons une solution intéressante proposée par Fabrice Moudjieu [1] concernant la modélisation des excès de zéros dans le cadre des Tree Pólya Splitting. Enfin, nous discutons d'ouvertures possibles pour traiter les valeurs extrêmes des mélanges Poisson et des modèles Pólya Splitting. En particulier, nous proposons une approche prometteuse qui combine les résultats du Chapitre 2 avec les relations que les modèles Splitting entretiennent avec les copules.

4.1 Représentations graphiques du modèle Tree Pólya Splitting

Comme présenté au Chapitre 3, la distribution Tree Pólya Splitting permet d'obtenir diverses relations de dépendance. Par exemple, on démontre que si le vecteur $\mathbf{Y}$ est distribué selon la distribution Tree Pólya Splitting présenté à la Figure 3.6, certaines paires de variables univariées sont indépendantes alors que d'autres sont négativement corrélées. Il peut être utile de représenter ces relations grâce à un modèle graphique probabiliste. Ces modèles permettent notamment d'interpréter certaines relations en terme de causalité. Il s'avère que certains cas particuliers de la Tree Pólya Splitting peuvent être représentés par un graphe mixte nommé graphe ancestral maximal.

Les modèles graphiques ancestraux proposés par Richardson et Spirtes [2002] sont une généralisation des graphes orientés acycliques utilisés dans les réseaux Bayésiens [Laurit-

1. Équipe GAMBAS (ANR-18-CE02-0025)

zen, 1996]. Zhang [2008] justifie cette généralisation par la nécessité de représenter des relations de causalité avec la présence de facteurs confondants. Ce modèle est donc un outil essentiel pour l'analyse causale [c.f. Richardson et Spirtes, 2003]. Afin d'utiliser ces graphes pour les modèles Tree Pólya Splitting, les notations et définitions utilisées par Richardson et Spirtes [2002] et Zhang [2008] sont présentées ci-dessous.

Notons par $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ un graphe avec sommets $\mathcal{V}$ et arêtes $\mathcal{E}$. Le graphe $\mathcal{G}$ est dit *orienté mixte* si les arêtes sont soit des flèches ($\leftarrow$, $\rightarrow$) ou des flèches doubles ($\leftrightarrow$). Pour deux sommets $X, Y \in \mathcal{V}$, X est un sommet *parent* de Y si $X \rightarrow Y$. Si $X \leftrightarrow Y$, alors les sommets (X, Y) sont *conjoints*. Un *chemin* de X vers Y dans le graphe $\mathcal{G}$ est l'ensemble de sommets $(X_k)_{k=0}^{K} \subseteq \mathcal{V}$ tels que $K \geq 1$, $X_0 = X$, $X_K = Y$ et il y a une arête entre X_k et X_{k+1} pour tout $k \in \{0, \ldots, K-1\}$. Ce chemin est dit *orienté* si X_k est le parent de X_{k+1} pour tout $k \in \{0, \ldots, K-1\}$. Le sommet X est un *ancêtre* de Y s'il existe un chemin orienté de X vers Y ou si $X = Y$. Pour X un ancêtre de Y, il existe un *cycle orienté* ou un *cycle semi-orienté* dans $\mathcal{G}$ si $Y \rightarrow X$ ou $Y \leftrightarrow X$ respectivement. Pour un chemin quelconque, un sommet non-terminal X_k représente une *intersection* si $\rightarrow X_k \leftarrow$, $\leftrightarrow X_k \leftrightarrow$, $\rightarrow X_k \leftrightarrow$ ou $\leftrightarrow X_k \leftarrow$. Finalement, il est nécessaire de définir un critère qui permet de séparer les sommets dans $\mathcal{G}$. Richardson et Spirtes [2002] proposent le critère de m-séparation.

Definition 4.1 (m-séparation)**.** Un chemin de X vers Y dans un graphe orienté mixte $\mathcal{G}$ est *actif* par rapport à un ensemble de sommets $\mathbf{Z}$ (peut-être vide), où $X, Y \notin \mathbf{Z}$, si les conditions suivantes sont vérifiées :

1. Tout sommet qui n'est pas une intersection dans le chemin n'est pas un élément de $\mathbf{Z}$;
2. Toute intersection dans le chemin est un ancêtre d'au moins un élément de $\mathbf{Z}$.

X et Y sont *m-séparés* par $\mathbf{Z}$ s'il n'existe aucun chemin actif entre X et Y par rapport à $\mathbf{Z}$. Par convention, si $\mathbf{Z} = \varnothing$, on dit que X et Y sont simplement m-séparés. Deux ensembles disjoints de sommets $\mathbf{X}$ et $\mathbf{Y}$ sont m-séparés par $\mathbf{Z}$ si tous les sommets de $\mathbf{X}$

sont m-séparés de tous les sommets de $\mathbf{Y}$ par rapport à $\mathbf{Z}$.

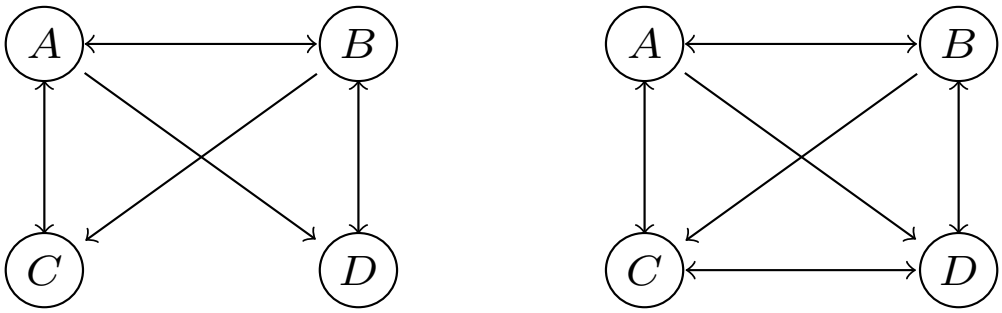
En combinant le critère de m-séparation aux définitions du graphe orienté mixte, Richardson et Spirtes [2002] définissent le graphe ancestral maximal.

Definition 4.2 (Richardson et Spirtes [2002])**.** Un graphe orienté mixte $\mathcal{G}$ est dit *ancestral maximal* (MAG) s'il est

1. Ancestral : le graphe ne contient aucun cycle orienté ou semi-orienté ;
2. Maximal : pour toute paire de sommets (X, Y) non adjacents dans $\mathcal{G}$, c'est-à-dire aucune arête entre X et Y, il existe un ensemble $\mathbf{Z}$ (peut-être vide) tel que X et Y sont m-séparés par rapport à $\mathbf{Z}$.

Afin de bien comprendre les définitions, étudions deux exemples proposés par Zhang [2008] et présentés à la Figure 4.1. Pour la figure (a), A et B sont des ancêtres de D et C respectivement. De plus, le graphe ne contient aucun cycle orienté ou semi-orienté. Il est donc ancestral. Par définition, le chemin $C \leftrightarrow A \leftrightarrow B \leftrightarrow D$ contient deux intersections : A et B. Ce chemin est actif par rapport à $\mathbf{Z} = (A, B)$ puisque les deux intersections sont des ancêtres de $\mathbf{Z}$. Pour le même ensemble $\mathbf{Z}$, le chemin $C \leftarrow B \leftrightarrow D$ est m-séparé puisque B n'est pas une intersection dans ce cas. Cependant, ce chemin devient actif lorsque $\mathbf{Z} = A$ ou $\mathbf{Z} = \varnothing$. De même, le chemin $C \leftrightarrow A \rightarrow D$ est actif pour $\mathbf{Z} = B$ ou $\mathbf{Z} = \varnothing$. Ainsi, le graphe (a) n'est pas maximal car il n'existe aucun ensemble $\mathbf{Z}$ qui m-sépare les sommets non adjacents C et D. Si on ajoute l'arête $C \leftrightarrow D$, alors le graphe (b) est bien maximal puisqu'il n'existe aucune paire de sommets non adjacents. Finalement, puisqu'il n'y a aucun cycle orienté ou semi-orienté, on obtient un MAG.

Comme mentionné précédemment, les graphes ancestraux généralisent les graphes orientés acycliques (DAG). En effet, si on retire les flèches doubles, les définitions de cycles et ancêtres restent similaires. Dans ce cas, le critère de m-séparation se ramène au critère



(a) (b)

FIGURE 4.1 – (a) Un graphe ancestral ; (b) Un graphe ancestral maximal

de d-séparation introduit par Pearl [1988]. De plus, il est démontré que les DAG sont toujours maximaux. Ainsi, les MAG correspondent exactement aux DAG lorsqu'on retire les flèches doubles. Pour plus de détails concernant les DAG, ou les modèles graphiques en général, voir Pearl [1988], Lauritzen [1996] ou Koller et Friedman [2009].

Finalement, afin d'incorporer la théorie des probabilités à celle des graphes, il est important de comprendre comment il est possible de représenter une distribution jointe avec un graphe. On rappelle que le critère de m-séparation permet de bloquer tous les chemins possibles entre deux ensembles de sommets disjoints dans un graphe MAG. Cette séparation s'apparente à l'indépendance conditionnelle où deux ensembles de variables aléatoires disjointes n'interagissent plus conditionnellement à un autre ensemble de variables aléatoires $\mathbf{Z}$. Dans ce contexte, on réfère au critère de séparation à la *propriété de Markov* de $\mathcal{G}$. Richardson et Spirtes [2002] définissent la propriété de Markov pour les MAG à l'aide du critère de m-séparation (Définition 4.1). Soit $\mathbf{Y} = (Y_1, \ldots, Y_J)$ un vecteur aléatoire correspondant à une mesure de probabilité $\mathbb{P}$ quelconque. Le graphe $\mathcal{G}$ est une *carte parfaite* de la mesure $\mathbb{P}$ si chaque sommet est associé à une marginale Y_j et toutes les indépendances conditionnelles de $\mathbb{P}$ sont décrites dans $\mathcal{G}$ à l'aide du critère de séparation. Précisément pour les MAG, $Y_i \perp\!\!\!\perp Y_j$ conditionnellement à $\mathbf{Z} \subset \mathbf{Y}$ si et seulement si Y_i et Y_j sont m-séparés par $\mathbf{Z}$ dans $\mathcal{G}$. Nous sommes enfin en mesure de

présenter certaines distributions Tree Pólya Splitting où leurs graphes MAG sont des cartes parfaites.

Exemple 1

Débutons par un exemple proposé par Peyhardi [2023]. Supposons pour $\mathbf{Y} = (Y_1, Y_2, Y_3)$ que $|\mathbf{Y}| \sim \mathcal{NB}(\theta_1 + \theta_2, p)$, où $\theta_1, \theta_2 \in \mathbb{R}_+$ et $p \in (0,1)$. De plus, supposons que l'arbre de partition $\mathfrak{T}$ est tel que ses nœuds internes sont donnés par $\mathfrak{I} = (\{1,2,3\}, \{1,2\})$. Pour le premier nœud, on suppose une multinomiale $\mathcal{M}_{\Delta_n}(\pi, 1-\pi)$ et, pour le deuxième nœud, une Dirichlet-multinomiale $\mathcal{DM}_{\Delta_n}(\theta_1, \theta_2)$. Cette distribution est donc représentée par l'arbre à la Figure 4.2. À l'aide du Théorème 1.21, les distributions de $Y_1 + Y_2$ et Y_3 sont données par

$$\begin{aligned} Y_1 + Y_2 &\sim \mathcal{B}_n(\pi) \underset{n}{\wedge} \mathcal{NB}(\theta_1 + \theta_2, p) \\ Y_3 &\sim \mathcal{B}_n(1-\pi) \underset{n}{\wedge} \mathcal{NB}(\theta_1 + \theta_2, p). \end{aligned}$$

De plus, grâce à Rao [1965] et le Théorème 6 de Peyhardi *et al.* [2021], les distributions de $Y_1 + Y_2$ et Y_3 sont des binomiales négatives. En effet, ces distributions sont équivalentes à

$$\begin{aligned} Y_1 + Y_2 &\sim \mathcal{NB}\left(\theta_1 + \theta_2, \frac{p\pi}{1 - p(1-\pi)}\right) \\ Y_3 &\sim \mathcal{NB}\left(\theta_1 + \theta_2, \frac{p(1-\pi)}{1 - p\pi}\right). \end{aligned}$$

Grâce au Théorème 1.22, la distribution de $Y_1 + Y_2$ permet d'obtenir l'indépendance au deuxième nœud. Précisément, $Y_1 \perp\!\!\!\perp Y_2$ et

$$Y_j \sim \mathcal{NB}\left(\theta_j, \frac{p\pi}{1 - p(1-\pi)}\right) \text{ pour } j = 1, 2.$$

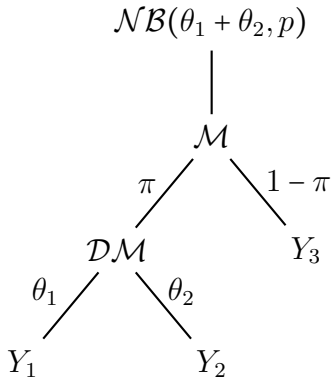


FIGURE 4.2 – Tree Pólya Splitting (Exemple 1)

Puisque la binomiale négative est surdispersée, on a $\mathrm{Cov}(Y_1+Y_2,Y_3)>0$. Ainsi, Y_1 et Y_2 sont corrélées positivement à Y_3 par la Proposition 3.8. Finalement, par une simple manipulation de la fonction de masse, on a le résultat suivant.

Théorème 4.1. Soit $\mathbf{Y}$ le vecteur aléatoire distribué selon la Tree Pólya Splitting à la Figure 4.3. On a les distributions conditionnelles suivantes :

$$Y_3 \mid (Y_1,Y_2) \sim \mathcal{NB}(\theta_1+\theta_2+Y_1+Y_2, p(1-\pi))$$
$$(Y_1,Y_2) \mid Y_3 \sim \mathcal{DM}_{\Delta_n}(\theta_1,\theta_2) \underset{n}{\wedge} \mathcal{NB}(\theta_1+\theta_2+Y_3, p\pi).$$

Démonstration. Il suffit d'étudier les distributions marginales du vecteur (Y_1+Y_2,Y_3) au premier nœud. Puisque $(Y_1+Y_2,Y_3) \sim \mathcal{M}_{\Delta_n}(\pi,1-\pi) \underset{n}{\wedge} \mathcal{NB}(\theta_1+\theta_2,p)$ et $Y_3 \sim \mathcal{NB}\left(\theta_1+\theta_2, \frac{p(1-\pi)}{1-p\pi}\right)$, on a

$$\begin{aligned}
\mathbb{P}(Y_1+Y_2=n \mid Y_3=y_3) &= \frac{\mathbb{P}(Y_1+Y_2=n, Y_3=y_3)}{\mathbb{P}(Y_3=y_3)} \\
&= \frac{\left[(n+y_3)!\frac{\pi^n}{n!}\frac{(1-\pi)^{y_3}}{y_3!}\right]\left[\frac{(\theta_1+\theta_2)_{n+y_3}}{(n+y_3)!}p^{n+y_3}(1-p)^{\theta_1+\theta_2}\right]}{\frac{(\theta_1+\theta_2)_{y_3}}{y_3!}\left(\frac{p(1-\pi)}{1-p\pi}\right)^{y_3}\left(\frac{1-p}{1-p\pi}\right)^{\theta_1+\theta_2}} \\
&= \frac{(\theta_1+\theta_2+y_3)_n}{n!}(p\pi)^n(1-p\pi)^{\theta_1+\theta_2+y_3}.
\end{aligned}$$

Un calcul similaire permet d'obtenir Y_3 conditionnellement à (Y_1,Y_2). □

Ainsi, en combinant les Théorèmes 1.22 et 4.1, Y_1 et Y_2 sont indépendantes conditionnellement à Y_3 si et seulement si $Y_3 = 0$. Lorsque $Y_3 > 0$, Y_1 et Y_2 sont nécessairement corrélées négativement. En effet, comme présenté à la Section 3.2.3, puisque $\mathcal{NB}(\theta_1 + \theta_2 + Y_3, p\pi)$ est surdispersée et ses moments factoriels sont tels que $\mu_1^2/(\mu_2 - \mu_1^2) = \theta_1 + \theta_2 + Y_3 > \theta_1 + \theta_2$, on peut conclure. Grâce aux dépendances établies, on peut représenter notre premier exemple par le MAG en Figure 4.3. En particulier, on obtient simplement un DAG.

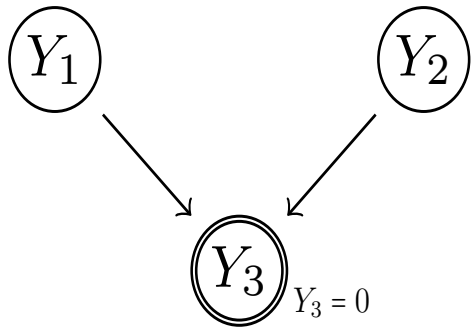


FIGURE 4.3 – MAG avec une exception au critère m-séparation pour le nœud Y_3 (Exemple 1)

Par le critère de m-séparation, Y_1 et Y_2 sont m-séparés et le chemin $Y_1 \to Y_3 \leftarrow Y_2$ devient actif par rapport à Y_3. Ce modèle Tree Pólya Splitting satisfait la première condition de Markov puisque Y_1 et Y_2 sont indépendantes. Cependant, $Y_1 \perp\!\!\!\perp Y_2$ sachant que $Y_3 = 0$ et donc le critère de m-séparation n'est pas vérifié pour ce cas. On doit alors indiquer dans le MAG qu'il y a une exception à la règle pour le sommet Y_3 (voir *context-specific* dans Koller et Friedman [2009]). À la Figure 4.3, on présente cette exception à l'aide d'un cercle double. On note que cette exception n'est pas présentée dans l'exemple de Peyhardi [2023].

Pour certains jeux de données, cet exemple de Tree Pólya Splitting peut avoir une interprétation intéressante. Supposons que $\mathbf{Y}$ représente un ensemble de proies (Y_1, Y_2) et un prédateur Y_3. On pourrait, par exemple, imaginer Y_3 comme étant une population de renards et (Y_1, Y_2) une population jointe d'écureuils et de lapins. Supposons aussi que

les deux familles de proies n'interagissent pas entre elles, c'est-à-dire qu'elles ne sont pas hostiles l'une envers l'autre. Cette hypothèse correspond à l'indépendance entre Y_1 et Y_2. Si la population de proies augmente, il est raisonnable de croire que la population de renards augmente également puisqu'ils auront accès à plus de nourriture. Parce que les covariances $\text{Cov}(Y_1, Y_3)$ et $\text{Cov}(Y_2, Y_3)$ sont positives, ce modèle est approprié pour une telle hypothèse. Enfin, si nous connaissons le nombre de prédateurs dans la région, cela devrait nous informer sur les deux populations de proies. Lorsqu'il n'y a aucun prédateur, c'est-à-dire $Y_3 = 0$, alors les proies n'interagissent toujours pas entre elles. Cependant, si $Y_3 > 0$, il devrait y avoir une corrélation négative entre les deux proies. En effet, si l'une des proies a une population réduite, cela implique que les prédateurs chassent principalement cette espèce, offrant l'opportunité à l'autre population de proies de croître. Ces deux situations sont adéquatement représentées par $(Y_1, Y_2) \mid Y_3$ dans notre modèle.

Exemple 2

Comme deuxième exemple, utilisons une structure similaire, mais cette fois la première division est remplacée par une multinomiale $\mathcal{M}_{\Delta_n}(\pi_{1,2}, \pi_3, \pi_4)$ (Voir Figure 4.4). En utilisant la même analogie proies/prédateurs, on remarquera que cette modification revient à ajouter un autre prédateur Y_4 qui interagit avec Y_3.

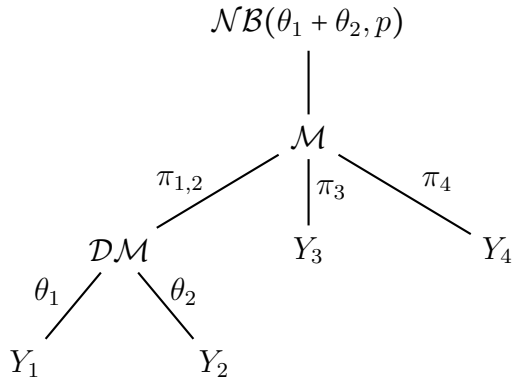


FIGURE 4.4 – Tree Pólya Splitting (Exemple 2)

À l'aide d'un argument similaire que celui du Théorème 4.1, on peut démontrer que les seules indépendances possibles sont (i) $Y_1 \perp\!\!\!\perp Y_2$ et (ii) $Y_1 \perp\!\!\!\perp Y_2$ conditionnellement à (Y_3, Y_4) si et seulement si $Y_3 = Y_4 = 0$. Cette dernière indépendance conditionnelle est également satisfaite avec seulement $Y_3 = 0$ ou $Y_4 = 0$. Grâce au critère de m-séparation, cet ensemble d'indépendances est représenté par le MAG en Figure 4.5. Comme dans notre premier exemple, il est nécessaire d'indiquer des sommets d'exception pour Y_3 et Y_4, mais également une arête d'exception pour les conjoints $Y_3 \leftrightarrow Y_4$. On présente cette dernière par une double flèche épaisse.

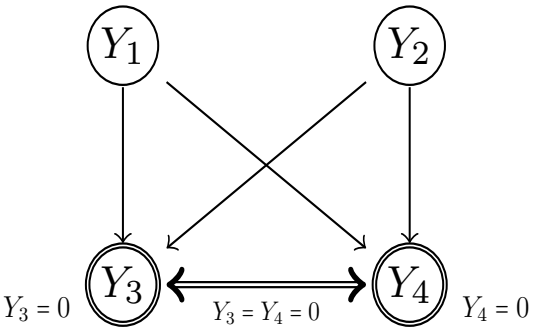


FIGURE 4.5 – MAG avec des exceptions au critère m-séparation pour les nœuds (Y_3, Y_4) et l'arête $Y_3 \leftrightarrow Y_4$ (Exemple 2)

Dans cet exemple, les covariances sont similaires à celles du premier exemple, avec l'interaction supplémentaire entre Y_3 et Y_4. Encore une fois, grâce à la surdispersion de la binomiale négative, $\text{Cov}(Y_3, Y_4) > 0$. Par conséquent, si ces deux variables aléatoires représentent des prédateurs, elles devraient avoir une corrélation positive dans ce modèle.

Exemple 3

Pour notre dernier exemple, généralisons le modèle précédent en divisant le vecteur (Y_3, Y_4) par une Dirichlet-multinomiale $\mathcal{DM}_{\Delta_n}(\theta_3, \theta_4)$ et en utilisant une binomiale né-

gative $\mathcal{NB}(\alpha, p)$ de paramètres $\alpha > 0$ et $p \in (0,1)$. L'objectif de cet exemple est d'étudier l'impact de α sur la représentation MAG du modèle Tree Pólya Splitting présentée à la Figure 4.6. Tout d'abord, supposons que $\alpha \neq \theta_1 + \theta_2$ et $\alpha \neq \theta_3 + \theta_4$. De plus, supposons $\theta_1 + \theta_2 - \alpha \notin \mathbb{N}$, de même pour (θ_3, θ_4). Encore une fois, par les mêmes calculs qu'au Théorème 4.1, on obtient

— $(Y_1, Y_2) \sim \mathcal{DM}_{\Delta_n}(\theta_1, \theta_2) \underset{n}{\wedge} \mathcal{NB}\left(\alpha, \frac{p\pi}{1-p(1-\pi)}\right)$;
— $(Y_1, Y_2) \mid (Y_3, Y_4) \sim \mathcal{DM}_{\Delta_n}(\theta_1, \theta_2) \underset{n}{\wedge} \mathcal{NB}\left(\alpha + Y_3 + Y_4, p\pi\right)$;
— $(Y_3, Y_4) \sim \mathcal{DM}_{\Delta_n}(\theta_3, \theta_4) \underset{n}{\wedge} \mathcal{NB}\left(\alpha, \frac{p(1-\pi)}{1-p\pi}\right)$;
— $(Y_3, Y_4) \mid (Y_1, Y_2) \sim \mathcal{DM}_{\Delta_n}(\theta_3, \theta_4) \underset{n}{\wedge} \mathcal{NB}\left(\alpha + Y_1 + Y_2, p(1-\pi)\right)$.

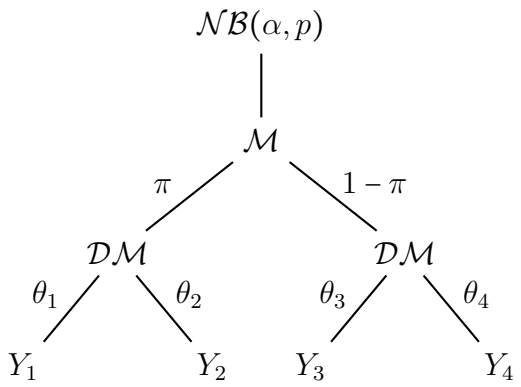


FIGURE 4.6 – Tree Pólya Splitting (Exemple 3)

Par la Proposition 3.8, la covariance au premier nœud est positive. De plus, par le Théorème 1.22 et nos hypothèses sur α, Y_1 est dépendante de Y_2 et Y_3 est dépendante de Y_4. Ainsi, aucune indépendance n'est possible pour toute paire de $\mathbf{Y}$. Pour les indépendances conditionnelles, les distributions de $(Y_1, Y_2)\big|(Y_3, Y_4)$ et $(Y_3, Y_4)\big|(Y_1, Y_2)$ sont de type Pólya Splitting possédant des corrélations non-nulles. De plus, nous pouvons démontrer le résultat suivant.

Théorème 4.2. Soit $\mathbf{Y}$ le vecteur aléatoire distribué selon la Tree Pólya Splitting à la Figure 4.6. Alors $(Y_1, Y_2) \mid Y_3 \sim \mathcal{DM}_{\Delta_n}(\theta_1, \theta_2) \underset{n}{\wedge} \tilde{\mathcal{L}}(\alpha, p, \pi, \theta_3, \theta_4, Y_3)$ où $\tilde{\mathcal{L}}$ est une

distribution univariée avec fonction de masse

$$\mathbb{P}(Y_1+Y_2=n\mid Y_3)=\frac{(\alpha+Y_3)_n}{n!}(p\pi)^n(1-p\pi)^{\alpha+Y_3}\frac{{}_2F_1\left[\begin{smallmatrix}\alpha+Y_3+n,\ \theta_4\\ \theta_3+\theta_4+Y_3\end{smallmatrix};p(1-\pi)\right]}{{}_2F_1\left[\begin{smallmatrix}\alpha+Y_3,\ \theta_4\\ \theta_3+\theta_4+Y_3\end{smallmatrix};\frac{p(1-\pi)}{1-p\pi}\right]}. \tag{4.1}$$

Démonstration. La probabilité conditionnelle est donnée par

$$\mathbb{P}(Y_1+Y_2=n\mid Y_3)=\frac{\mathbb{P}(Y_1+Y_2=n)}{\mathbb{P}(Y_3=y_3)}\sum_{y_4=0}^{\infty}\mathbb{P}(Y_3=y_3,Y_4=y_4\mid Y_1+Y_2=n) \tag{4.2}$$

où, par le Théorème 4.1,

$$(Y_3,Y_4)\big|Y_1+Y_2\sim\mathcal{DM}_{\Delta_n}(\theta_3,\theta_4)\underset{n}{\wedge}\mathcal{NB}(\alpha+Y_1+Y_2,p(1-\pi)).$$

De plus, $Y_3\sim\mathcal{BB}_n(\theta_3,\theta_4)\underset{n}{\wedge}\mathcal{NB}\left(\alpha,\frac{p(1-\pi)}{1-p\pi}\right)$ et $Y_1+Y_2\sim\mathcal{NB}\left(\alpha,\frac{p\pi}{1-p(1-\pi)}\right)$. Ces trois distributions possèdent les fonctions de masse suivantes :

$$\mathbb{P}(Y_1+Y_2=n)=\frac{(\alpha)_n}{n!}\left(\frac{p\pi}{1-p(1-\pi)}\right)^n\left(\frac{1-p}{1-p(1-\pi)}\right)^{\alpha},$$

$$\mathbb{P}(Y_3=y_3)=\left(\frac{1-p}{1-p\pi}\right)^{\alpha}\frac{(\alpha)_{y_3}(\theta_3)_{y_3}}{(\theta_3+\theta_4)_{y_3}y_3!}\left(\frac{p(1-\pi)}{1-p\pi}\right)^{y_3}{}_2F_1\left[\begin{smallmatrix}\alpha+y_3,\ \theta_4\\ \theta_3+\theta_4+y_3\end{smallmatrix};\frac{p(1-\pi)}{1-p\pi}\right],$$

$$\begin{aligned}\mathbb{P}(Y_3=y_3,Y_4=y_4\mid Y_1+Y_2=n)=&\frac{(\alpha+n)_{y_3}(\theta_3)_{y_3}}{(\theta_3+\theta_4)_{y_3}y_3!}(p(1-\pi))^{y_3}(1-p(1-\pi))^{\alpha+n}\\&\cdot\left[\frac{(\alpha+n+y_3)_{y_4}(\theta_4)_{y_4}}{(\theta_3+\theta_4+y_3)_{y_4}y_4!}(p(1-\pi))^{y_4}\right].\end{aligned}$$

L'insertion de ces dernières dans (4.2) permet d'obtenir le résultat. □

Les fonctions de masse pour $(Y_1,Y_2)\big|Y_4$, $(Y_3,Y_4)\big|Y_1$ et $(Y_3,Y_4)\big|Y_2$ sont similaires à celle en (4.1). Encore une fois, par le Théorème 1.22, aucune indépendance conditionnelle n'existe. Par conséquent, le MAG associé à la distribution Tree Pólya Splitting est donné par le graphe entièrement connecté présenté à la Figure 4.7.

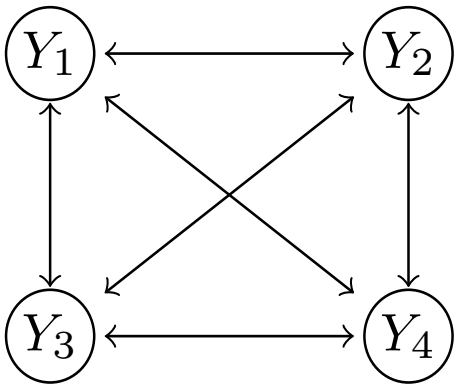


FIGURE 4.7 – MAG pour $\alpha \neq \theta_1 + \theta_2$ et $\alpha \neq \theta_3 + \theta_4$ (Exemple 3)

Changeons maintenant les hypothèses précédentes avec $\alpha = \theta_1 + \theta_2$. Dans ce cas, par un argument similaire, on a (i) $Y_1 \perp\!\!\!\perp Y_2$, et (ii) $Y_1 \perp\!\!\!\perp Y_2$ conditionnellement à (Y_3, Y_4) si et seulement si $Y_3 = Y_4 = 0$. Cependant, en utilisant les Théorèmes 1.22 et 4.2, $Y_1 \not\perp\!\!\!\perp Y_2$ conditionnellement à Y_3 ou Y_4. Pour cette situation, le MAG est donné par la Figure 4.8 avec seulement l'arête d'exception entre Y_3 et Y_4. Si $\alpha = \theta_3 + \theta_4$, le MAG possède une structure similaire où les paires (Y_1, Y_2) et (Y_3, Y_4) sont échangées.

En utilisant l'interprétation proies/prédateurs pour la Figure 4.8, nous avons toujours une indépendance entre les proies (Y_1, Y_2) et une indépendance conditionnelle si les deux prédateurs (Y_3, Y_4) ne sont pas présents. Cependant, les proies restent dépendantes conditionnellement à un seul prédateur, ce qui diffère du modèle à la Figure 4.5. Pour ce qui est de la corrélation entre les prédateurs, celle-ci varie selon la valeur de α. Si $\alpha > \theta_3 + \theta_4$ ou $\alpha < \theta_3 + \theta_4$, alors (Y_3, Y_4) possèdent une corrélation négative ou positive respectivement. De plus, si $\alpha < \theta_3 + \theta_4$ et les valeurs de (Y_1, Y_2) sont connues, il possible que la corrélation devienne négative. En effet, puisque $(Y_3, Y_4) \mid (Y_1, Y_2) \sim \mathcal{DM}_{\Delta_n}(\theta_3, \theta_4) \underset{n}{\wedge} \mathcal{NB}(\alpha + Y_1 + Y_2, p(1-\pi))$, il existe des valeurs de Y_1 et Y_2 telles que $\alpha + Y_1 + Y_2 > \theta_3 + \theta_4$. La corrélation entre les prédateurs est, dans ce cas, positive. Une interprétation de cette propriété peut être la suivante : sans aucune connaissance de la population de proies, les deux prédateurs interagissent amicalement entre eux. Cepen-

dant, si nous connaissons le nombre de proies dans la région et qu'il est suffisamment élevé, cela peut impliquer que les prédateurs deviennent plus hostiles les uns envers les autres, car, sinon, la population de proies diminuerait.

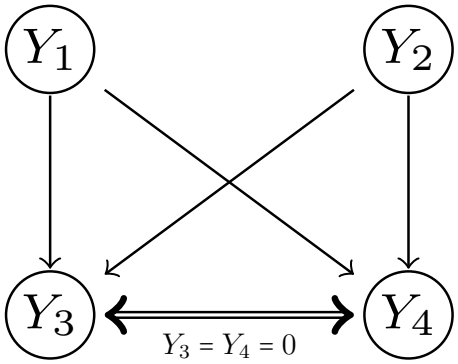


FIGURE 4.8 – MAG pour $\alpha = \theta_1 + \theta_2$, $\alpha \neq \theta_3 + \theta_4$ avec une exception au critère m-séparation pour l'arête $Y_3 \leftrightarrow Y_4$ (Exemple 3)

Finalement, supposons que la binomiale négative est telle que $\alpha = \theta_1 + \theta_2 = \theta_3 + \theta_4$. Dans ce cas, (i) $Y_1 \perp\!\!\!\perp Y_2$, (ii) $Y_3 \perp\!\!\!\perp Y_4$, (iii) $Y_1 \perp\!\!\!\perp Y_2$ conditionnellement à (Y_3, Y_4) si et seulement si $Y_3 = Y_4 = 0$ et (iv) $Y_3 \perp\!\!\!\perp Y_4$ conditionnellement à (Y_1, Y_2) si et seulement si $Y_1 = Y_2 = 0$. Autrement, grâce au Théorème 4.2, (Y_1, Y_2) sont dépendantes conditionnellement à Y_3 ou Y_4 et (Y_3, Y_4) sont dépendantes conditionnellement à Y_1 ou Y_2. Ainsi, le graphe MAG présenté à la Figure 4.9 est tel qu'il y a une flèche double entre chaque paire de $\mathbf{Y}$, à l'exception des paires (Y_1, Y_2) et (Y_3, Y_4). On indique également les exceptions au critère m-séparation lorsqu'on conditionne par rapport à (Y_1, Y_2) ou (Y_3, Y_4).

4.2 Modèle Tree Pólya Splitting avec excès de zéros

Le traitement des valeurs nulles est essentiel pour plusieurs domaines scientifiques. Que ce soit en écologie [cf. Wenger et Freeman, 2008; Blasco-Moreno *et al.*, 2019], en microbiologie [cf. Xu *et al.*, 2015; Tang et Chen, 2018] ou en actuariat [cf. Boucher *et al.*, 2007,

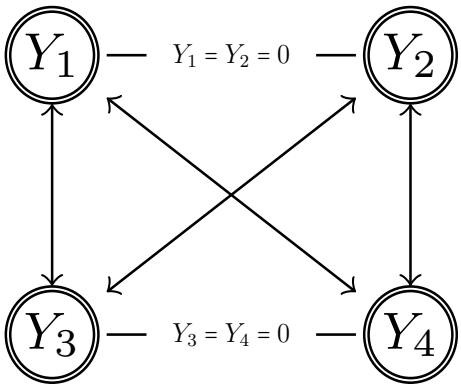


FIGURE 4.9 – MAG pour $\alpha = \theta_1 + \theta_2 = \theta_3 + \theta_4$ avec des exceptions au critère m-séparation pour les paires (Y_1, Y_2) et (Y_3, Y_4) (Exemple 3)

2009], l'incorporation des excès de zéros permet d'améliorer la qualité des modèles. Dans le cadre univarié, les modèles d'excès de zéros Poisson [Lambert, 1992], binomiale négative [Greene, 1994], Poisson généralisée [Gupta *et al.*, 1996] ou Conway-Maxwell-Poisson [Sellers et Raim, 2016] sont tous des exemples de modèles largement utilisés. Chacun de ces modèles utilise le principe suivant. Supposons $Z \sim \text{Ber}(\pi)$, la distribution de Bernoulli et $Y \sim F_Y$ une variable aléatoire univariée discrète indépendante de Z. Le modèle des excès de zéros est alors la distribution du produit ZY. Cette approche peut être appliquée au cadre multivarié.

Soit $\mathbf{Y} = (Y_1, \ldots, Y_J)$ un vecteur aléatoire discret, indépendant de Z, distribué selon une loi $F_{\mathbf{Y}}$. Une simple généralisation du modèle univarié est d'étudier la distribution de $Z\mathbf{Y}$. Cette approche est utilisée, par exemple, lorsque les Y_j sont des variables aléatoires indépendantes Poisson [Liu et Tian, 2015] ou Conway-Maxwell-Poisson [Santana *et al.*, 2022]. Malheureusement, le seul type d'excès de zéros considéré dans ces modèles est le vecteur complet, c'est-à-dire $(0, \ldots, 0)$. Pour remédier à cette contrainte, Liu et Tian [2015] proposent que les Y_j soient des variables aléatoires indépendantes du modèle d'excès de zéros Poisson. Ainsi, chaque combinaison d'excès de zéros est considérée. Plus précisément, cette distribution multivariée augmente les probabilités d'observer les

vecteurs $(0, Y_2, \ldots, Y_J)$, $(Y_1, 0, \ldots, Y_J)$, $(0, 0, \ldots, Y_J), \ldots, (0, 0, \ldots, 0)$. Néanmoins, il n'est pas réaliste de supposer que les marginales soient indépendantes.

À des fins d'applications en microbiologie, Tang et Chen [2018] présentent un modèle d'excès de zéros de la Dirichlet-multinomiale généralisée. On rappelle que la Dirichet-multinomiale généralisée est un cas particulier du modèle Tree Pólya Splitting où l'arbre de partition $\mathfrak{T}$ est un arbre binaire en cascade, chaque nœud interne $\mathfrak{I}$ étant associé à une bêta-binomiale, c'est-à-dire une Dirichlet-multinomiale de dimension 2, et la somme à la racine est une valeur fixe. Afin d'adapter cette distribution, Tang et Chen [2018] proposent d'utiliser à chaque nœud une distribution bêta-binomiale avec excès de zéros. Plus précisément, pour $Z \sim \text{Ber}(\pi)$ et $Y \sim \mathcal{BB}_n(\alpha, \beta)$ indépendantes, la distribution à chaque nœud correspond à celle du produit ZY. Notons cette distribution par $\mathcal{ZI}-\mathcal{DM}_{\Delta_n}(\alpha, \beta; \pi)$. Visuellement, on peut interpréter ce modèle comme un "gonflement" des zéros à chaque branche gauche de l'arbre $\mathfrak{T}$. Autrement dit, cette Tree Pólya Splitting particulière augmente les chances d'envoyer une valeur nulle à chaque feuille de gauche. Voir Figure 4.10 pour un exemple.

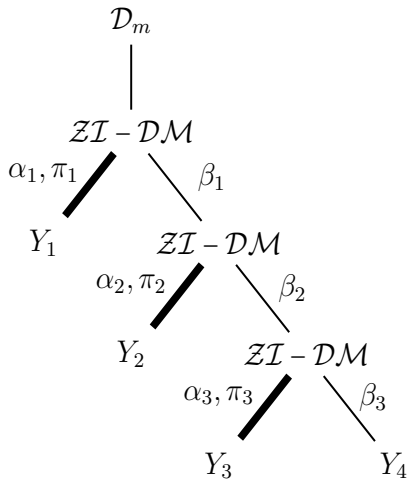


FIGURE 4.10 – Modèle d'excès de zéros de la Dirichlet-multinomiale généralisée

Puisque ce type de modèle utilise une structure d'arbre de partition similaire à la Tree

Pólya Splitting, il semble naturel qu'on puisse étendre cette approche. Pour ce faire, un collègue doctorant travaillant également dans le cadre de l'ANR GAMBAS (ANR-18-CE02-0025), Fabrice Moudjieu, propose de généraliser ce modèle en utilisant la structure d'un arbre binaire quelconque. Plus précisément, il utilise une distribution Tree Pólya Splitting où chaque nœud interne est une bêta-binomiale avec excès de zéros soit à gauche, soit à droite. Ainsi, il généralise non seulement l'arbre utilisé par Tang et Chen [2018], mais aussi le principe de "gonflement" des branches. Grâce à cette approche, il est en mesure de modéliser des données d'abondance d'arbres et d'incorporer des covariables à ses analyses. Nous pouvons donc constater que les modèles Tree Pólya Splitting sont polyvalents et leur utilisation est intuitive.

4.3 Valeurs extrêmes pour les mélanges Poisson

Comme indiqué brièvement en conclusion du Chapitre 2, il est possible qu'une distribution discrète possède un domaine d'attraction lorsque ses paramètres varient selon la taille d'échantillon. Par exemple, Anderson *et al.* [1997] démontrent que la loi de Poisson avec paramètre λ_n peut être dans le domaine d'attraction de Gumbel lorsque $\lim_{n\to\infty} \lambda_n = \infty$ selon un certain taux de croissance. Évidemment, la distribution ne possède aucun domaine d'attraction si le paramètre est fixé. La preuve d'Anderson *et al.* [1997] est fondée sur l'heuristique suivante : puisque la distribution de Poisson peut être approchée par une normale lorsque $\lambda_n \to \infty$ et que cette dernière est dans $\mathcal{D}_0$, il est raisonnable de croire que la loi de Poisson est également dans ce domaine d'attraction à la limite. Ainsi, Anderson *et al.* [1997] caractérisent le taux de croissance de λ_n et obtiennent des suites normalisantes des maxima de variables Poisson. De façon similaire, Nadarajah et Mitov [2002] démontrent que si $\lim_{n\to\infty} p_n = 1$ avec un certain taux de croissance, alors il existe des suites normalisantes telles que la binomiale négative $\mathcal{NB}(\alpha, p_n)$ est dans le domaine

d'attraction de Gumbel.

Par le résultat de Willmot [1990] et le Corollaire 2.1, il est possible de connecter plusieurs mélanges Poisson aux deux résultats précédents. En effet, Willmot [1990] démontre que si λ a comme densité

$$f(x) \sim C(x)x^{\alpha-1}e^{-\beta x}, \text{ lorsque } x \to \infty, \tag{4.3}$$

où $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}_+$ et $C(x)$ est une fonction à variation lente localement bornée, alors la fonction de masse du mélange Poisson se comporte de manière similaire à la binomiale négative. Dans le même ordre d'idée, le Corollaire 2.1 démontre que si $\lambda \in \mathcal{D}_-$, alors le mélange Poisson ressemble asymptotiquement à une distribution Poisson. Ces équivalences asymptotiques nous amènent à penser que les suites normalisantes de Anderson [1970] et Nadarajah et Mitov [2002] pourraient être également utilisées pour ces mélanges Poisson lorsque certains de leurs paramètres varient selon la taille d'échantillon. Selon des résultats préliminaires, cela semble être le cas pour le mélange Poisson où λ est distribuée selon une densité gaussienne inverse généralisée [Sichel, 1974]. En effet, nous pouvons démontrer que cette densité satisfait la relation d'équivalence (4.3) et, pour un cas particulier, que les suites normalisantes de Nadarajah et Mitov [2002] peuvent être adaptées afin que le mélange Poisson soit dans le domaine d'attraction de Gumbel. Une piste intéressante serait d'explorer davantage cette application des résultats de Willmot [1990] et Nadarajah et Mitov [2002] pour plusieurs distributions de mélange satisfaisant (4.3). De manière similaire, nous souhaitons explorer ces techniques en combinant le Corollaire 2.1 aux suites normalisantes d'Anderson *et al.* [1997].

4.4 Valeurs extrêmes pour les modèles Splitting

Dans cette thèse, nous nous sommes intéressés à la théorie des valeurs extrêmes discrètes univariées, en mettant l'accent sur les mélanges Poisson. Cependant, qu'en est-il des valeurs extrêmes discrètes multivariées ? Plus précisément, qu'en est-il des valeurs extrêmes pour les modèles Pólya Splitting ? Pour tenter d'élucider une réponse, revenons à un cas particulier présenté à la Section 1.4. Soit $\mathcal{M}(\boldsymbol{\psi})$ une distribution réelle positive, supposons que

$$\mathbf{Y} \sim \mathcal{DM}_{\Delta_n} \underset{n}{\wedge} \left[\mathcal{P}(\lambda) \underset{\lambda}{\wedge} \mathcal{M}(\boldsymbol{\psi}) \right].$$

Nous avons démontré que ce modèle Splitting est équivalent au mélange Poisson multivarié

$$\mathbf{Y} \sim \left[\prod_{j=1}^{J} \mathcal{P}(R_j) \right] \underset{\mathbf{R}}{\wedge} \mathcal{R}(\boldsymbol{\psi}, \boldsymbol{\theta}),$$

où $\mathbf{R} = (R_1, \ldots, R_J)$ est le vecteur aléatoire du produit $\lambda\boldsymbol{\pi}$ tel que $\boldsymbol{\pi} \sim \mathcal{D}_{\Delta}(\boldsymbol{\theta})$, la loi de Dirichlet. Dans un premier temps, nous constatons que les distributions marginales sont elles-mêmes des mélanges Poisson. Par conséquent, nous pouvons appliquer directement nos résultats du Chapitre 2 pour décrire leurs comportements asymptotiques. Cependant, est-il possible d'analyser la structure de dépendance de ces extrêmes ?

Rappelons-nous que le vecteur $\mathbf{R}$ est caractérisé par des copules de survie particulières. Plus précisément, nous savons que $\mathbf{R}$ possède une copule de survie archimédienne lorsque $\boldsymbol{\theta} = \mathbf{1}$ [McNeil et Nešlehová, 2009], ou une copule de survie Liouville lorsque $\boldsymbol{\theta} \in \mathbb{R}_+^J$ [McNeil et Nešlehová, 2010]. Ces copules décrivent comment les marginales R_j interagissent entre elles, mais aussi la façon dont leurs extrêmes univariées interagissent. En effet, le théorème 5.2.3 de Galambos [1987] stipule que les valeurs extrêmes d'une distribution continue multivariée peuvent être analysées par les domaines d'attraction des marginales

et d'une copule dite extrême. Cette dernière, notée par $\mathcal{C}^*$, est caractérisée par la limite

$$\lim_{t\to\infty} \mathcal{C}^t(u_1^{1/t}, \ldots, u_J^{1/t}) = \mathcal{C}^*(u_1, \ldots, u_J), \tag{4.4}$$

où $u_1, \ldots, u_J \in [0,1]$ et $\mathcal{C}$ est la copule de la distribution multivariée. Dans notre cas, Larsson et Nešlehová [2011] et Belzile et Nešlehová [2017] ont caractérisé la copule extrême de $\mathbf{R}$ lorsque $\boldsymbol{\theta} = \mathbf{1}$ et $\boldsymbol{\theta} \in \mathbb{R}_+^J$ respectivement. Dans le cas des distributions discrètes multivariées, Feidt *et al.* [2010] démontrent que les conditions de Galambos [1987] sont toujours adéquates pour décrire les extrêmes multivariés. Ainsi, en utilisant la connexion entre notre modèle Splitting et le vecteur $\mathbf{R}$, nous conjecturons que la copule extrême de ce dernier caractérise les interactions extrêmes du vecteur $\mathbf{Y}$. Cette approche semble être une piste prometteuse pour de nouvelles contributions sur les valeurs extrêmes discrètes multivariées et requiert une étude plus approfondie.

Bibliographie

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

S. Anders et W. Huber : Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

C. W. Anderson : Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability*, 7(1):99–113, 1970.

C. W. Anderson, S. G. Coles et J. Hüsler : Maxima of Poisson-like variables and related triangular arrays. *The Annals of Applied Probability*, 7(4):953–971, 1997.

A. A. Balkema et L. De Haan : On R. Von Mises' condition for the domain of attraction of $\exp(-e^{-x})$. *The Annals of Mathematical Statistics*, 43(4):1352–1354, 1972.

A. A. Balkema et L. De Haan : Residual life time at great age. *The Annals of Probability*, 2(5):792 – 804, 1974.

B. Bartoszewicz : Modelling the claim count with Poisson regression and negative binomial regression. *In Innovations in Classification, Data Science, and Information Systems*, pages 103–110. Springer Berlin Heidelberg, 2005.

L. R. Belzile et J. Nešlehová : Extremal attractors of Liouville copulas. *Journal of Multivariate Analysis*, 160:68–92, 2017.

P. Bhagwat : Models and Statistical Inference for Multivariate Count Data. Mémoire de maîtrise, Indian Institute of Science Education and Research Pune, 2019.

N.H. Bingham, C.M. Goldie et J.L. Teugels : *Regular Variation*. Cambridge University Press, 1987.

A. Blasco-Moreno, M. Pérez-Casany, P. Puig, M. Morante et E Castells : What does a zero mean ? Understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, 10(7):949–959, 2019.

L. N. Bol'shev : On a characterization of the Poisson distribution and its statistical applications. *Theory of Probability & Its Applications*, 10(3):446–456, 1965.

J.-P. Boucher, M. Denuit et M. Guillen : Risk classification for claim counts : A comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal*, 11(4):110–131, 2007.

J.-P. Boucher, M. Denuit et M. Guillen : Number of accidents or number of claims ? An approach with zero-inflated Poisson models for panel data. *Journal of Risk and Insurance*, 76(4):821–846, 2009.

J. P. Boucher, M. Denuit et M. Guillén : Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Variance*, 2(1):135–162, 2008.

K. M. Briggs, L. Song et T. Prellberg : A note on the distribution of the maximum of a set of Poisson random variables. *arXiv :0903.4373*, 2009.

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

M. G. Bulmer : On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, 30:101, 1974.

J. T. Campbell : The Poisson correlation function. *Proceedings of the Edinburgh Mathematical Society*, 4(1):18–26, 1934.

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

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

D. B. H. Cline : Convolution tails, product tails and domains of attraction. *Probability Theory and Related Fields*, 72:529–557, 1986.

J. F. Coeurjolly et J. Rousseau Trépanier : The median of a jittered Poisson distribution. *Metrika*, 83:837–851, 2020.

S. Coles : *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

R. J. Connor et 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 et 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 et 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.

K.T. Fang, S. Kotz et K.W. Ng : *Symmetric Multivariate and Related Distributions*. Monographs on Statistics and Applied Probability. Springer US, 2013.

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

W. Feller : On a general class of "contagious" distributions. *The Annals of Mathematical Statistics*, 14(4):389 – 400, 1943.

R. A. Fisher et L. H. C. Tippett : Limiting forms of the frequency distribution of the largest or smallest member of a sample. *In Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press, 1928.

J. Galambos : *The Asymptotic Theory of Extreme Order Statistics*. R.E. Krieger Pub, 1987.

L. Gardes et S. Girard : Estimation de quantiles extrêmes pour les lois à queue de type Weibull : une synthèse bibliographique. *Journal de la Société Française de Statistique*, 154(2):98–118, 2013.

C. Genest et J. Nešlehová : A primer on copulas for count data. *ASTIN Bulletin*, 37 (2):475–515, 2007.

B. V. Gnedenko : Sur la distribution limite du terme maximum d'une serie aléatoire. *Annals of Mathematics*, 44:423, 1943.

W. H. Greene : Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *NYU working paper no. EC-94-10*, 1994.

M. Greenwood et G.U. Yule : An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, 83:255–279, 1920.

Q.F. Gronau, H. Singmann et E. Wagenmakers : Bridgesampling : An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10):1–29, 2020.

P. L. Gupta, R. C. Gupta et R. C. Tripathi : Analysis of zero-adjusted count data. *Computational Statistics & Data Analysis*, 23(2):207–218, 1996.

R. D. Gupta et D. S.P. Richards : Multivariate Liouville distributions. *Journal of Multivariate Analysis*, 23(2):233–256, 1987.

L. De Haan : *On Regular Variation and Its Application to the Weak Convergence of Sample Extremes*. Mathematisch Centrum, 1970.

A. Hitz, R. Davis et G. Samorodnitsky : Discrete extremes. *arXiv :1707.05033*, 2017.

D. I. Inouye, E. Yang, G. I. Allen et P. Ravikumar : A review of multivariate distributions for count data derived from the Poisson distribution. *WIREs Computational Statistics*, 9(3):e1398, 2017.

J. O. Irwin : The generalized Waring distribution applied to accident theory. *Journal of the Royal Statistical Society. Series A (General)*, 131(2):205–225, 1968.

K. G. Janardan et 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, A. W. Kemp et S. Kotz : *Univariate discrete distributions*, volume 444. John Wiley & Sons, 2005.

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

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

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

A. Kaul, S. Mandal, O. Davidov et S. D. Peddada : Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, 8:283205, 2017.

AC Kimber : A note on Poisson maxima. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 63(4):551–552, 1983.

C. Kleiber et S. Kotz : *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley, 2003.

D. Koller et N. Friedman : *Probabilistic graphical models : principles and techniques*. MIT press, 2009.

D. Lambert : Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

M. Larsson et J. Nešlehová : Extremal behavior of Archimedean copulas. *Advances in Applied Probability*, 43(1):195–216, 2011.

S. L Lauritzen : *Graphical models*, volume 17. Clarendon Press, 1996.

M. R. Leadbetter, G. Lindgren et H. Rootzén : *Extremes and related properties of random sequences and processes*. Springer Science & Business Media, 2012.

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

A. J. McNeil et J. Nešlehová : Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. *The Annals of Statistics*, 37(5B):3059 – 3097, 2009.

A. J. McNeil et J. Nešlehová : From Archimedean to Liouville copulas. *Journal of Multivariate Analysis*, 101(8):1772–1790, 2010.

X. L. Meng et W. H. Wong : Simulating ratios of normalizing constants via a simple identity : A theoretical exploration. *Statistica Sinica*, 6(4):831–860, 1996.

S. Nadarajah et K. Mitov : Asymptotics of maxima of discrete random variables. *Extremes*, 5(3):287, 2002.

T. Nagler : A generic approach to nonparametric function estimation with mixed data. *Statistics and Probability Letters*, 137:326–330, 2018.

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

O. Ovaskainen et N. Abrego : *Joint species distribution modelling : with applications in R*. Cambridge University Press, 2020.

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

A. Panagiotelis, C. Czado et H. Joe : Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072, 2012.

G. P. Patil et M. V. Ratnaparkhi : Problems of damaged random variables and related characterizations. *A Modern Course on Statistical Distributions in Scientific Work*, pages 255–270, 1975.

J. Pearl : *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan kaufmann, 1988.

R. Perline : Mixed Poisson distributions tail equivalent to their mixing distributions. *Statistics and Probability Letters*, 38(3):229–233, 1998.

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

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

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

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

J. Pickands : Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

R Core Team : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

B. R. Rao et K. G. Janardan : The use of the generalized Markov-Pólya distribution as a random damage model and its identifiability. *Sankhyā : The Indian Journal of Statistics, Series A (1961-2002)*, 46(3):458–462, 1984.

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

S. I. Resnick : *Extreme Values, Regular Variation and Point Processes*. Springer, 1987.

S. I. Resnick : *Heavy-tail phenomena : probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

T. Richardson et P. Spirtes : Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962 – 1030, 2002.

T. S. Richardson et P. Spirtes : Causal inference via ancestral graph models. *Oxford Statistical Science Series*, pages 83–105, 2003.

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

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

J. Schulz, C. Genest et M. Mesfioui : A multivariate Poisson model based on comonotonic shocks. *International Statistical Review*, 89(2):323–348, 2021.

K. F. Sellers et A. Raim : A flexible zero-inflated model to address data dispersion. *Computational Statistics & Data Analysis*, 99:68–80, 2016.

M. Shaked : On mixtures from exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):192–198, 1980.

T. Shimura : Discretization of distributions in the maximum domain of attraction. *Extremes*, 15(3):299–317, 2012.

H. S. Sichel : On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society. Series A (General)*, 137(1):25–34, 1974.

M. Sklar : Fonctions de répartition à N dimensions et leurs marges. *Annales de l'ISUP*, VIII(3):229–231, 1959.

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

Stan Development Team : RStan : the R interface to Stan, 2020. URL `http://mc-stan.org/`. R package version 2.21.2.

A. G. Stephenson : evd : Extreme value distributions. *R News*, 2(2), 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

U.R. Subramanya : On max domains of attraction of univariate p-max stable laws. *Statistics & Probability Letters*, 19(4):271–279, 1994.

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

H. Teicher : On the multivariate Poisson distribution. *Scandinavian Actuarial Journal*, 1954(1):1–9, 1954.

P Usseglio-Polatera et 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, J. Peyhardi, G. Toulemonde, É. Marchand et F. Mortier : Tree Pólya Splitting distributions for multivariate count data. *⟨hal-04563659⟩*, 2024.

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

M. Vuilleumier : Sur le comportement asymptotique des transformations linéaires des suites. *Math Z*, 23:126–139, 1967.

T. Wang et 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 et F. K.C. Hui : So many variables : Joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12):766–779, 2015.

S. J. Wenger et M. C. Freeman : Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, 89(10):2953–2959, 2008.

R. E. Williamson : Multiply monotone functions and their Laplace transforms. *Duke Mathematical Journal*, 23(2):189 – 207, 1956.

G. E. Willmot : Asymptotic tail behaviour of Poisson mixtures with applications. *Advances in Applied Probability*, 22(1):147–159, 1990.

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.

L. Xu, A. D. Paterson, W. Turpin et W. Xu : Assessment and selection of competing models for zero-inflated microbiome data. *PloS one*, 10(7):e0129606, 2015.

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

J. Zhang : Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.

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