Analyse du Λ-coalescent : renouer avec ses racines.

SALA Raphaël, MUGISHA Axcel, GARCIA Hugo, COLIN Thibault

Novembre 2025

1 Introduction

1.1 Fondements du Λ-coalescent

La théorie de la coalescence modélise le phénomène par lequel des individus d'une population partagent un ancêtre commun. Nous souhaitons étudier rétrospectivement leur évolution.

Historiquement, le modèle de Wright-Fisher étudie une population de taille finie N où les individus d'une générations coalescent de manière uniforme entre eux dans la génération précédente [Fis30]. Ensuite, le modèle de Kingman [Kin82] est le modèle limite de Wright-Fisher où l'on s'intéresse à $n < N$ lignées et en considérant $N \to +\infty$. Ce cadre asymptotique permet de simplifier grandement l'étude du phénomène de coalescence. Le modèle peut à présent être décrit comme un processus de Markov.

En 1999, Pitman et Sagitov généralisent le modèle de Kingman en autorisant la coalescence simultanée de plusieurs lignées. Des individus peuvent engendrer une proportion non négligeable de la population. Afin de définir un modèle, nous supposons raisonnablement que les lignées coalescent aléatoirement et indépendamment de leur histoire passée, c'est-à-dire en supposant l'absence de mémoire (propriété de Markov), que toutes les lignées ont les mêmes chances de coalescer entre elles que l'on appelle l'échangeabilité et enfin que nous ayons l'absence de collisions multiples signifiant qu'à tout instant donné, il ne peut y avoir qu'un seul événement de fusion en un même ancêtre.

Théorème 1 (Pitman-Sagitov [Pit99, Sag99])**.** *Il existe un processus de Markov, $(N_t)_{t\geq 0}$, appelé Λ-coalescent, échangeable à collisions multiples simples si et seulement s'il existe une mesure finie Λ sur $[0,1]$ telle que, lorsqu'on a b lignées, pour tout $2 \leq k \leq b$ le taux auquel chaque k-uplet fixé de lignées fusionne vaut,*

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\,\Lambda(dx)$$

Nous ne définissons par formellement les conditions ici et donnons encore moins une preuve car cela est au-delà du cadre de ce rapport. Ce résultat montre que la dynamique de $(N_t)_{t\geq 0}$, indiquant le nombre de lignées à l'instant t, est entièrement caractérisée par une mesure finie. Sans perte de généralité nous considérons pour la suite une mesure de probabilité, Λ sur $[0,1]$. Partant de b lignées, le taux d'une k-coalescence ($2 \leq k \leq b$) est $r_{b,k} := \binom{b}{k}\lambda_{b,k}$. Le taux de sortie de l'état b est la somme des taux donc

$$\lambda_b = \sum_{k=2}^{b} r_{b,k} = \int_0^1 S_b(x)\,\Lambda(dx), \quad S_b(x) := \sum_{k=2}^{b}\binom{b}{k}x^{k-2}(1-x)^{b-k} = \frac{1-(1-x)^b - bx(1-x)^{b-1}}{x^2} \tag{1}$$

D'après le lemme des réveils, à chaque événement de coalescence on passe de b à $b-k+1$ lignées avec probabilité,

$$\forall b \geq k \geq 2, \quad p_{b,k} := \frac{r_{b,k}}{\sum_{k=2}^{b} r_{b,k}} = \frac{\binom{b}{k}\lambda_{b,k}}{\lambda_b}$$

Ainsi, le squelette du processus est une chaîne de Markov décroissante sur $[\![1,n]\!]$, commençant en n et absorbée presque sûrement en 1.

1.2 Exemple (Kingman)

Intéressons-nous au modèle de Kingman en guise d'introduction. On pose $\Lambda = \delta_0$. Pour $2 \leq k \leq b$,

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\delta_0(x)dx = [x^{k-2}(1-x)^{b-k}]_{x=0} = \begin{cases} (1-0)^{b-2} = 1 & \text{si } k = 2 \\ 0^{k-2}(1-0)^{b-k} = 0 & \text{si } k > 2 \end{cases}$$

Les coalescences se font que par paires. Une caractéristique intéressante du Λ-coalescent est le TMRCA (Time to the Most Recent Common Ancestor), c'est-à-dire le plus petit temps tel que toutes les lignées ont fusionné en un ancêtre commun. Dans la suite du rapport, nous le notons

$$\tau_{\Lambda,n} := \inf\{t \geq 0, N_t = 1\} \mid \{N_0 = n\}$$

Lorsque le contexte est clair sur Λ ou n, ceux-ci seront omis afin de rendre la lecture plus agréable.

Lemme 1. *Soit Λ une mesure de probabilité sur $[0,1]$. Notons $H : b \in \mathbb{N}^* \longmapsto \mathbb{E}(\tau_{\Lambda,b})$. Alors $H(1) = 0$, $H(2) = 1$ et pour $b \geq 3$,*

$$H(b) = \frac{1}{\lambda_b} + \sum_{k=2}^{b-1} p_{b,k}\, H(b-k+1)$$

Démonstration. Soit Λ une mesure de probabilité sur $[0,1]$ dont nous omettons sa présence dans les notations. Pour $b = 1$, $N_t = 1$ donc $H(1) = 0$. Pour $b = 2$, le seul saut possible est de 2 vers 1 lignée avec taux $\lambda_2 = \binom{2}{2}\lambda_{2,2} = 1$, d'où $\tau_2 \sim \text{Exp}(1)$ et $H(2) = 1/1 = 1$.

Fixons $b \geq 3$. Définissons le temps de la première coalescence,

$$T_b^1 := \inf\{t \geq 0, N_t \neq b\} \mid \{N_0 = b\}$$

$(N_t)_{t\geq 0}$ est un processus de Markov avec un taux de saut λ_b, donc $T_b^1 \sim \text{Exp}(\lambda_b)$ et donc $\mathbb{E}(T_b^1) = \frac{1}{\lambda_b}$. De plus, si K est la taille de la fusion au temps T_b^1, alors $K \sim \sum_{k=2}^b p_{b,k}\delta_k$ et $N_{T_b^1} = b - K + 1$.

Considérons la filtration naturelle $(\mathcal{F}_t)_{t\geq 0}$ de $(N_t)_{t\geq 0}$. Par la propriété de Markov forte au temps T_b^1 et l'absence de mémoire,

$$\mathbb{E}(\tau_b - T_b^1 \mid \mathcal{F}_{T_b^1}) = \mathbb{E}(\tau_{N_{T_b^1}}) = H(N_{T_b^1})$$

Ainsi en conditionnant par $\mathcal{F}_{T_1}$,

$$H(b) = \mathbb{E}(\tau_b) = \mathbb{E}(T_b^1) + \mathbb{E}(\tau_b - T_b^1) = \frac{1}{\lambda_b} + \mathbb{E}\Big(\mathbb{E}(\tau_b - T_b^1 \mid \mathcal{F}_{T_b^1})\Big) = \frac{1}{\lambda_b} + \mathbb{E}(H(N_{T_b^1}))$$
$$= \frac{1}{\lambda_b} + \sum_{k=2}^{b} \mathbb{P}\Big(N_{T_b^1} = b-k+1\Big)H(b-k+1) = \frac{1}{\lambda_b} + \sum_{k=2}^{b} p_{b,k}H(b-k+1)$$

Or $H(1) = 0$, donc le terme $k = b$ s'annule. D'où le résultat. □

Pour b lignées observées, on a $\lambda_b = \sum_{k=2}^b \binom{b}{k}\lambda_{b,k} = \binom{b}{2}\lambda_{b,2} = \binom{b}{2}$ donc, d'après le Lemme 1, la taille moyenne d'un arbre pour le modèle de Kingman est donné par,

$$H(b) = \frac{1}{\lambda_b} + p_{b,2}\, H(b-1) = \frac{1}{\binom{b}{2}} + H(b-1)$$

Ainsi par récurrence,

$$H(b) = \sum_{k=2}^{b} \frac{1}{\binom{k}{2}} = \sum_{k=2}^{b} \frac{2}{k(k-1)} = \sum_{k=2}^{b} 2\Big(\frac{1}{k-1} - \frac{1}{k}\Big) = 2\Big(1 - \frac{1}{b}\Big) \tag{2}$$

Nous illustrons ce résultat par une simulation pour $n = 20$ lignées (figure 1). On observe bien que les fusions sont binaires. L'état 1 est absorbant et dans ce cas ci le TMRCA est supérieur à $\tau_{\delta_0,n}$. La grande

partie des coalescences se font dans les premiers instants. En référence au lemme des réveils, le nombre de paires possibles, et donc de réveils prêts à sonner et paramétrés de la même manière, sont plus nombreux au début qu'à la fin. L'arbre a donc l'aspect d'un acacia de la savane [1], dense et large en haut, fin en bas.

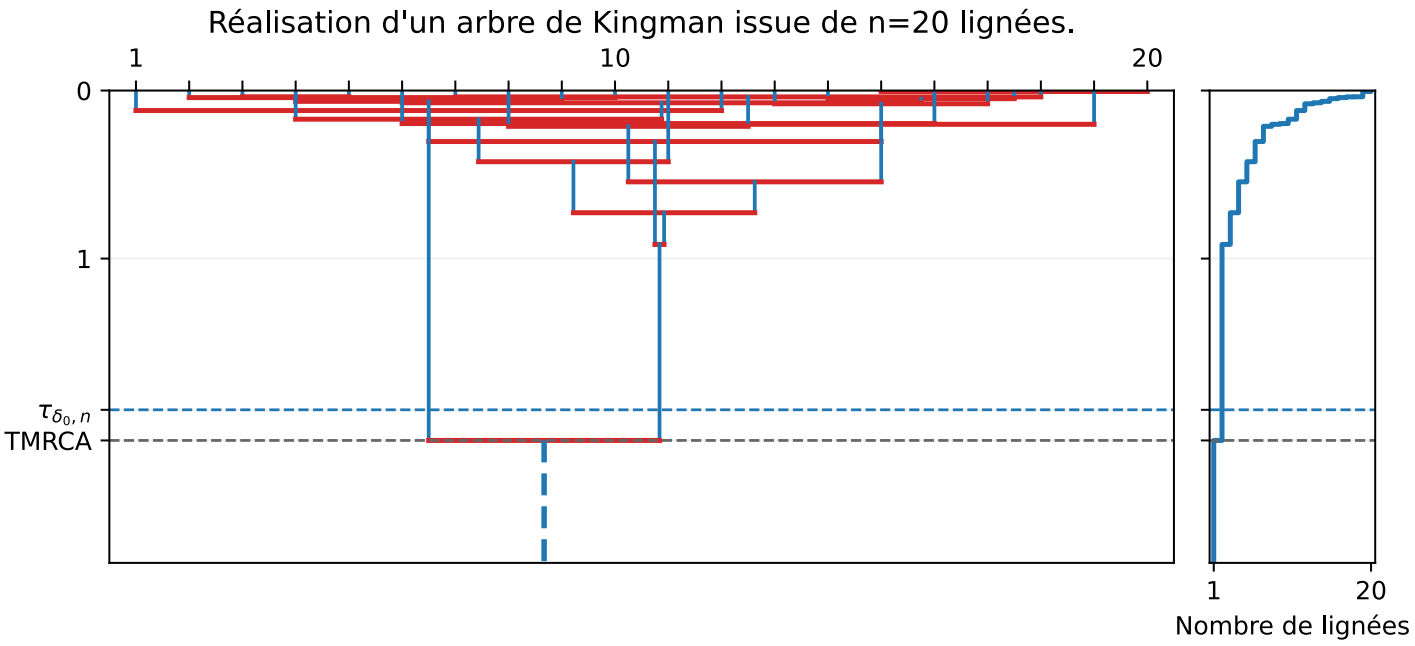

FIGURE 1 – Réalisation du coalescent de Kingman issu de $n = 20$ lignées. À gauche, représentation de l'arbre généalogique et mesure du TMRCA comparé à la valeur théorique $\tau_{\delta_0,n} = 2(1 - 1/20) = 1.9$. À droite, évolution du nombre de lignées N_t jusqu'à l'état absorbant $N_t = 1$ où nous atteignons un plateau constant pour $t > TMRCA$.

2 Analyse du TMRCA

2.1 Aux extrêmes de l'arbre.

Au vu du précédent exemple, on peut se demander l'influence de la mesure Λ sur le TMRCA. Intuitivement, ce temps moyen devrait diminuer lorsque la masse de Λ se rapproche de 1 puisqu'on autorise des coalescences multiples plus importantes. En première analyse on va étudier les deux cas extrêmes.

Proposition 1. *Soit n le nombre de lignées. Alors, pour toute mesure de probabilité Λ sur $[0, 1]$, on a,*

$$1 = \mathbb{E}(\tau_{\delta_1,n}) \leq \mathbb{E}(\tau_{\Lambda,n})$$

Démonstration. Prouvons l'égalité. Prenons $\Lambda = \delta_1$, nous avons $\lambda_{n,k} = \delta_{n,k}$ (symbole de Kronecker), donc $\lambda_n = \binom{n}{n}\lambda_{n,n} = 1$ donc $\tau_{\delta_1} \sim \text{Exp}(1)$ et donc $\mathbb{E}(\tau_{\delta_1}) = 1/1 = 1$.

Soit Λ une mesure de probabilité sur $[0, 1]$. Notons $H(b) := \mathbb{E}(\tau_{\Lambda,b})$.

Montrons par récurrence forte l'inégalité, c'est-à-dire $H(b) \geq 1$ pour $b \geq 2$. L'initialisation a été prouvée dans le lemme 1. Supposons l'inégalité vraie jusqu'à $b - 1$. Remarquons que $\lambda_{b,b} = \int_0^1 x^{b-2}\,\Lambda(dx) \leq \int_0^1 \Lambda(dx) = 1$,

$$H(b) = \frac{1}{\lambda_b} + \sum_{k=2}^{b-1} p_{b,k} H(b-k+1) \geq \frac{1}{\lambda_b} + \sum_{k=2}^{b-1} p_{b,k} = \frac{1}{\lambda_b} + 1 - p_{b,b} = 1 + \frac{1-\lambda_{b,b}}{\lambda_b} \geq 1$$

D'où le résultat. □

1. Référence au Vachellia tortilis, qui n'est plus considéré comme un acacia.

Cette idée de déplacer la masse de Λ vers 1 pour diminuer la moyenne du TMRCA est intuitive. Pour le problème inverse de maximisation du TMRCA nous souhaiterions déplacer la masse de Λ vers 0. C'est-à-dire prouver que le modèle de Kingman soit celui maximisant le temps moyen du TMRCA. Toutefois, voila une grande surprise : ce n'est pas le cas !

Proposition 2. *Il existe $n > 1$ et une mesure de probabilité Λ sur $[0, 1]$ telle que,*

$$\mathbb{E}(\tau_{\Lambda,n}) > \mathbb{E}(\tau_{\delta_0,n})$$

Démonstration. Soit $n = 8$, dans l'exemple de Kingman (voir sous-section 1.2), nous avons une formule explicite.

$$\mathbb{E}(\tau_{\delta_0}) = 2\left(1 - \frac{1}{8}\right) = \frac{14}{8} = 1.75$$

Soit $\Lambda = \delta_{1/4}$, alors d'après (1),

$$\lambda_{n,k} = \left(\frac{1}{4}\right)^{k-2}\left(\frac{3}{4}\right)^{n-k} \qquad \lambda_n = 16\left(1 - \left(\frac{3}{4}\right)^n - \frac{n}{4}\left(\frac{3}{4}\right)^{n-1}\right)$$

Ainsi, en calculant nous obtenons, toujours pour $n = 8$,

$$\mathbb{E}(\tau_{\delta_{1/4},n}) = \frac{1}{\lambda_n} + \sum_{k=2}^{n-1} \frac{\binom{n}{k}\lambda_{n,k}}{\lambda_n}\mathbb{E}(\tau_{\delta_{1/4},n-k+1}) = \frac{19954284839411683}{11337879079537330} \approx 1.7599662\ldots > 1.75$$

□

Nous conjecturons que le théorème peut être étendu pour tout $n > 6$. A notre connaissance l'étude ce phénomène n'est pas documenté pour n fini. Seul un article de [KLLS17] s'intéresse la croissance de $\sup_\Lambda \mathbb{E}(\tau_\Lambda)$ lorsque $n \to \infty$.

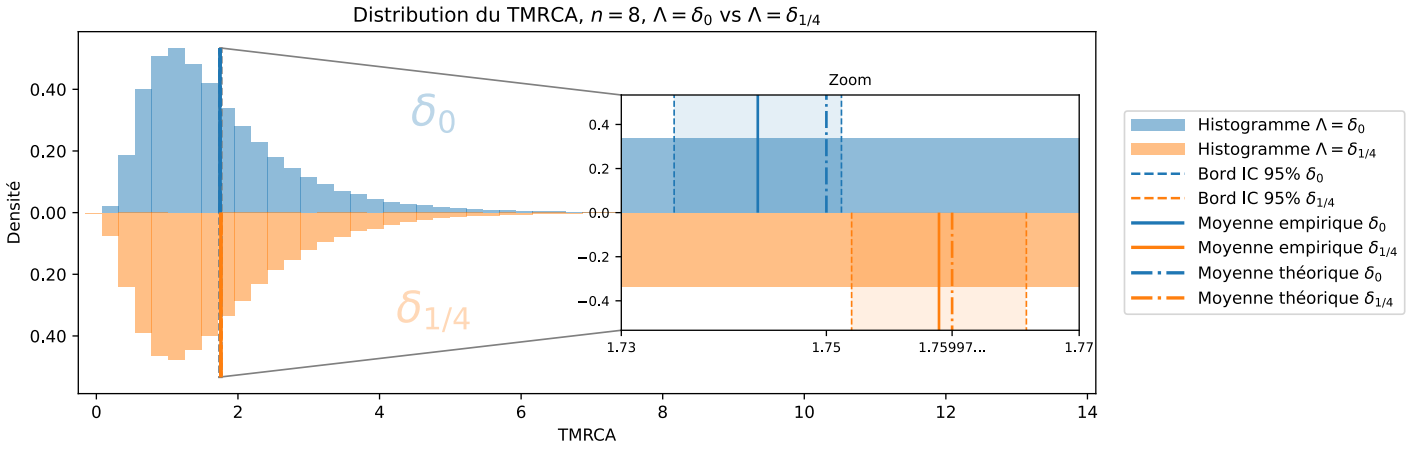

FIGURE 2 – Distribution empirique du TMRCA pour $n = 8$ de $\Lambda = \delta_0$ (Kingman, histogramme vers le haut en bleu) et $\Lambda = \delta_{1/4}$ (histogramme vers le bas en orange). Les bandes verticales pointillées délimitent les intervalles de confiance (IC) à 95% pour la moyenne de τ basés sur $M = 1e5$ simulations indépendantes. Les lignes pleines indiquent les moyennes empiriques et les lignes pointillées ("-.") les moyennes théoriques.

Dans la figure 2 nous utilisons des intervalles de confiances. Soit $(T_i)_{1\leq i\leq M}$ une collection de variables aléatoires i.i.d. de loi τ_Λ. Posons la moyenne et l'estimateur de la variance

$$\overline{T}_M := \frac{1}{M}\sum_{i=1}^{M} T_i \qquad s_M^2 := \frac{1}{M-1}\sum_{i=1}^{M}(T_i - \overline{T}_M)^2$$

Les intervalles de confiances sont construits génériquement. D'après le théorème central limite et le lemme de Slutsky,

$$\sqrt{M}\frac{\overline{T}_M - \mathbb{E}_\Lambda[\tau]}{\sqrt{s_M^2}} \xrightarrow[M\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Ainsi, un IC asymptotique de niveau $1-\alpha$ est

$$[\overline{T}_M \pm q_{1-\alpha/2}\frac{\sqrt{s_M^2}}{\sqrt{M}}]$$

Dans la figure 2 le zoom à droite montre les intervalles de confiance à 95% disjoints renforçant l'observation $\mathbb{E}(\tau_{\delta_0,8}) < \mathbb{E}(\tau_{\delta_{1/4},8})$.

2.2 Une forêt pas si grande

Un processus de Markov est entièrement déterminé par son générateur infinitésimal. Pour n lignées observées, celui d'un Λ-coalescent est la matrice triangulaire inférieure $Q \in \mathcal{M}_n(\mathbb{R})$ définie pour tout $1 \le b, i \le n$, par

$$Q_{b,i} = \begin{cases} r_{b,k} & \text{si } b \ge 2 \text{ et } i = b-k+1 \text{ pour } 2 \le k \le b \\ -\lambda_b & \text{si } b \ge 2 \text{ et } i = b \\ 0 & \text{sinon} \end{cases}$$

Le premier élément de sa diagonale, $Q_{1,1}$, est nul car l'état 1 est absorbant donc Q n'est pas inversible. En se restreignant à la sous-matrice $R = (Q_{i,j})_{2\le i,j\le n}$ la matrice devient inversible et nous pouvons exprimer la densité de τ. Posons $p_R(t) = (p_k(t))_{2\le k\le n}$ où $p_k : t \ge 0 \mapsto \mathbb{P}(N_t = k \mid N_0 = n)$. D'après la relation de Chapman-Kolmogorov, p_R vérifie pour tout $t \ge 0$

$$\begin{cases} p_R'(t) = p_R(t)R \\ p_R(0) = (0,\dots,0,1) \end{cases} \iff p_R(t) = (0,\dots,0,1)e^{tR}$$

Définissons la fonction de survie, $S : t \mapsto \mathbb{P}(\tau_n > t) = \mathbb{P}(N_t \ne 1 \mid N_0 = n) = \sum_{k=2}^n p_k(t) = p_R(t)\cdot\mathbf{1}$. Donc la densité de τ_n est donnée par,

$$f_\tau : t \mapsto d_t(1-S(t)) = -S'(t) = -p_R'(t)\cdot\mathbf{1} = -p_R(t)R\cdot\mathbf{1} = -(0,\dots,0,1)e^{tR}R\cdot\mathbf{1} \tag{3}$$

On remarque également que ce processus est défini par $(\lambda_{b,k})_{I_n}$ avec $I_n \coloneqq \{(b,k), 2 \le k \le b \le n\}$. Définissons pour $r \in [\![0, n-2]\!]$

$$m_r : \Lambda \mapsto \int_0^1 x^r \Lambda(dx)$$

En développant l'intégrande des taux de fusions, pour tout $(b,k) \in I_n$, il existe $A_n \in \mathcal{M}_{I_n,n-1}(\mathbb{R})$ tel que,

$$\lambda_{b,k} = \sum_{r=0}^{n-2} A_{(b,k),r} m_r(\Lambda)$$

Pour n fixé, Q est entièrement déterminée par $(m_r(\Lambda))_{0\le r\le n-2}$, l'espace des mesures de probabilité sur $[0,1]$ se réduit à une projection de dimension finie, $\mathbb{R}^{n-1}$, donc un espace bien plus petit. Autrement dit, une infinité de mesures différentes deviennent indiscernables pour un processus considéré.

Proposition 3. *Soit $n > 1$ et $\Lambda_0^\alpha \coloneqq \mathrm{Beta}(2-\alpha,\alpha)$ avec $\alpha \in]0,2[$, de densité*

$$w_\alpha : x \in [0,1] \longmapsto \frac{1}{B(2-\alpha,\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}$$

Soit $(J_n)_{n\ge0}$ les polynômes de Jacobi. Pour tout $n \ge 0$, $J_n = \sum_{k=0}^n \binom{n+\alpha-1}{n-k+1}\binom{n-\alpha+1}{k} x^{n-k}(x-1)^k$ est de degré n et orthogonal à tous les polynômes de degré inférieur à $n-1$ pour le produit scalaire [NEM94]

$$\langle f, g\rangle_\alpha = \int_0^1 f(x)g(x)w_\alpha(x)dx$$

On pose

$$M \coloneqq \sup_{x\in[0,1]} |J_{n-1}(x)| \in]0, +\infty[\qquad et \qquad \varepsilon_n \coloneqq \frac{1}{M}$$

Et on définit pour $0 < \varepsilon < \varepsilon_n$*, la mesure de probabilité* $\Lambda_\varepsilon^\alpha \neq \Lambda_0^\alpha$ *par sa densité*

$$f_\varepsilon^\alpha : x \in [0,1] \longmapsto \Big(1 + \varepsilon J_{n-1}(x)\Big)w_\alpha(x)$$

Alors, pour tout $0 < \varepsilon < \varepsilon_n$*,*

$$\tau_{\Lambda_\varepsilon^\alpha, n} \overset{\mathcal{L}}{=} \tau_{\Lambda_0^\alpha, n}$$

Démonstration. Soit $\varepsilon < 1/M$, montrons que f_ε^α est bien une densité. Sur $[0,1]$, $1 + \varepsilon J_{n-1} \geq 1 - \varepsilon M \geq 0$ donc par produit de termes positifs $f_\varepsilon^\alpha \geq 0$.

Puis, par orthogonalité de J_{n-1} avec la constante $1 \in \mathbb{R}_{n-1}[X]$,

$$\int_0^1 J_{n-1}(x)w_\alpha(x)dx = 0$$

D'où,

$$\int_0^1 f_\varepsilon^\alpha(x)dx = \int_0^1 w_\alpha(x)dx + \varepsilon\int_0^1 J_{n-1}(x)w_\alpha(x)dx = 1 + \varepsilon\cdot 0 = 1$$

Ainsi f_ε^α est bien une densité sur $[0,1]$.

Montrons à présent que les générateurs infinitésimaux de Λ_0^α et $\Lambda_\varepsilon^\alpha$ coïncident. Pour tout $r \in [\![0, n-2]\!]$, on a $X^r \in \mathbb{R}_r[X] \subset \mathbb{R}_{n-1}[X]$, ainsi par orthogonalité de J_{n-1},

$$\int_0^1 x^r J_{n-1}(x)w_\alpha(x)\,dx = 0$$

Ainsi,

$$m_r(\Lambda_\varepsilon^\alpha) = \int_0^1 x^r f_\varepsilon(x)\,dx = \int_0^1 x^r w_\alpha(x)\,dx + \varepsilon\int_0^1 x^r J_{n-1}(x)w_\alpha(x)\,dx = m_r(\Lambda_0^\alpha)$$

Les taux de fusions sont donc égaux entre ces mesures, d'où le résultat. □

Ainsi nous venons de construire une infinité de mesures différentes qui induisent le même processus de coalescence. On s'attendait à obtenir une infinité d'arbres généalogiques différents, mais ceux-ci sont identiques en loi.

Pour la construction nous utilisons les polynômes de Jacobi. Ce choix est motivé par la stucture de la mesure $\mathrm{Beta}(2-\alpha, \alpha)$ et du fait que les polynômes de Jacobi sont orthogonaux pour le poids de la forme $x^\beta(1-x)^\gamma$ sur $[0,1]$ avec $\beta, \gamma > -1$. Le poids w_α implique que $\beta = \alpha - 1, \gamma = 1 - \alpha$ et la condition sur β et γ impose que $\alpha \in]0, 2[$.

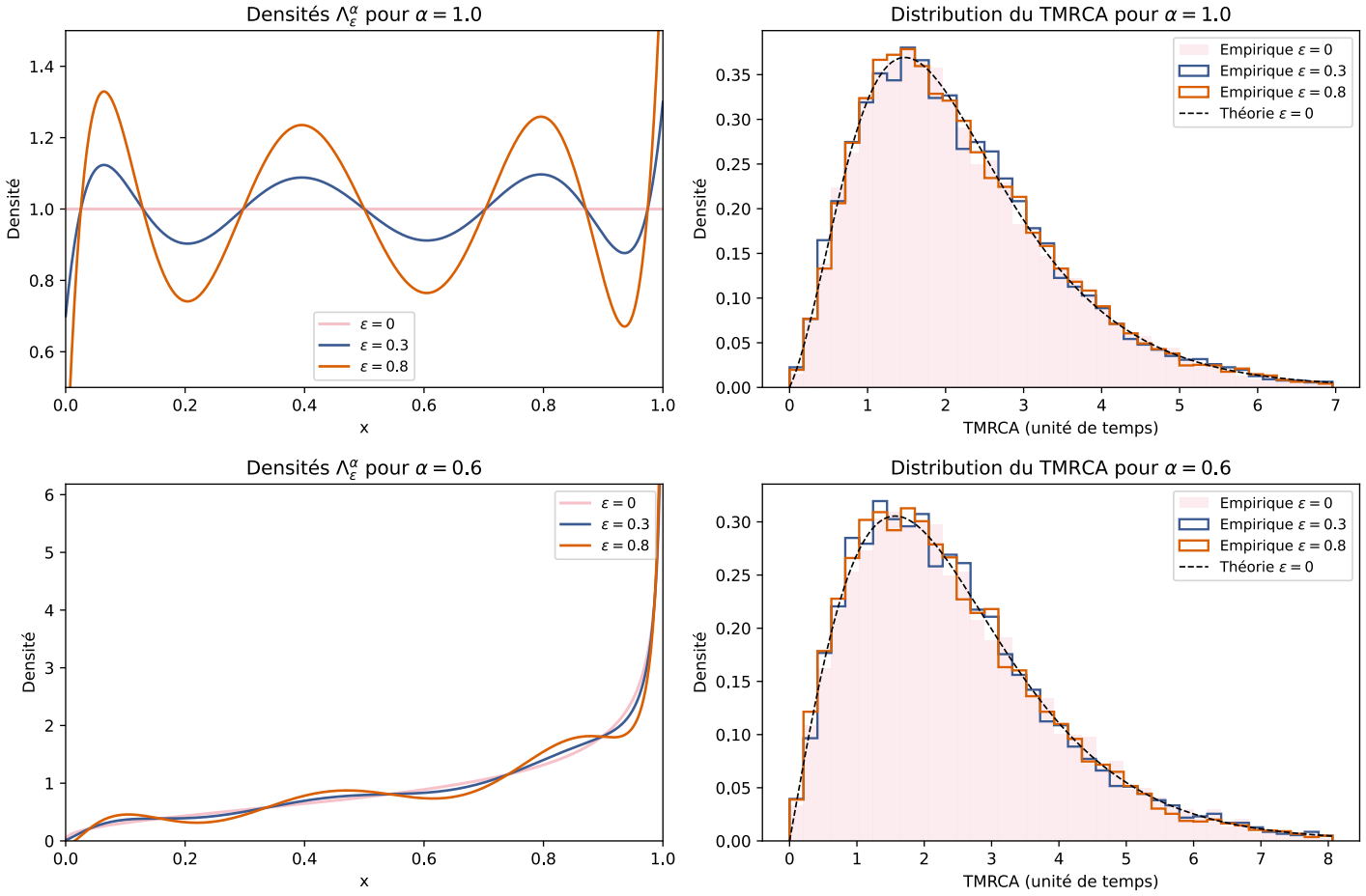

FIGURE 3 – (a) *Gauche* : Graphe des densités[2] $\Lambda_\varepsilon^\alpha(x) = (1 + \varepsilon J_{n-1}(x))w_\alpha(x)$ pour $n = 8$ et $\varepsilon \in \{0, 0.3, 0.8\}$, dans les cas $\alpha = 1$ (haut) et $\alpha = 0.6$ (bas). On voit que ε contrôle l'amplitude des oscillations autour de la densité de référence $\varepsilon = 0$. (b) *Droite* : distributions empiriques du TMRCA issues de 8000 simulations de n lignées, avec superposition de la densité théorique prise pour $\varepsilon = 0$ donnée par (3). Malgré des densités $\Lambda_\varepsilon^\alpha$ très différentes, la loi du TMRCA est identique à celle du modèle de référence.

La construction que nous proposons permet de jouer avec les densités selon 2 paramètres. L'allure générale est dictée par α tandis que ε intensifie l'amplitude des oscillations. Le choix de ε est contraint d'être inférieur à ε_n. D'après [Sze39], cette borne est assez restrictive,

$$\varepsilon_n = \Theta_{n\to\infty}\left(\frac{1}{n^{|\alpha-1|}}\right)$$

Dans la figure 3 sur les graphiques du bas avec $\alpha = 0.6$, nous avons pour $n = 8$ que $\varepsilon_n = 0.392062$. On remarque alors que la distribution empirique vérifie bien la théorie pour $\varepsilon = 0.2 < \varepsilon_n$ mais aussi pour $\varepsilon = 0.8 > \varepsilon_n$. La contrainte sur ε est seulement pour maintenir une densité positive.

Dans le cas $\alpha = 1$, graphiques du haut, nous obtenons la mesure uniforme et ε_n ne depend plus de n. Toutefois, le choix de ε est tout de même borné par 1 afin que notre densité reste positive.

2.3 Silence, ça pousse

Précedemment nous avons brièvement parlé de la mesure uniforme en prenant Λ_0^α avec $\alpha = 1$. Ce modèle est connu sous le nom de Bolthausen-Sznitman et décèle un résultat incontournable.

Théorème 2 (Goldschmidt & Martin [GM05]). *Soit $(N_t)_{t\geq 0}$ un Bolthausen-Sznitman coalescent. Alors,*

$$\tau_n - \log(\log(n)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{G}$$

2. Sauf pour $\alpha = 0.6$ et $\varepsilon = 0.8$ qui n'est pas une densité, comme expliqué dans les paragraphes qui suivent.

où $\mathcal{G}$ est la loi de Gumbel, de densité $x \mapsto e^{-x-e^{-x}}$.

La preuve est omise car elle dépasse le cadre de ce rapport. Toutefois, ce théorème renforce l'intuition qu'on a pu commencé à avoir à la Proposition 2 puisqu'on a que,

$$\lim_{n\to\infty} \mathbb{E}(\tau_n) = \lim_{n\to\infty} \log(\log(n)) + \mathbb{E}(\mathcal{G}) = \lim_{n\to\infty} \log(\log(n)) + \gamma = +\infty$$

où γ est la constante d'Euler-Mascheroni. En effet, comme l'illustre la figure 4, dès $n \approx 50$ on observe $\mathbb{E}(\tau_n) > 2$, surpassant la borne du modèle de Kingman (2). Ainsi, la croissance de la hauteur des arbres généalogiques pour le modèle de Bolthausen-Sznitman est extrêmement lente mais permet d'obtenir des arbres aussi grand que l'on souhaite en moyenne.

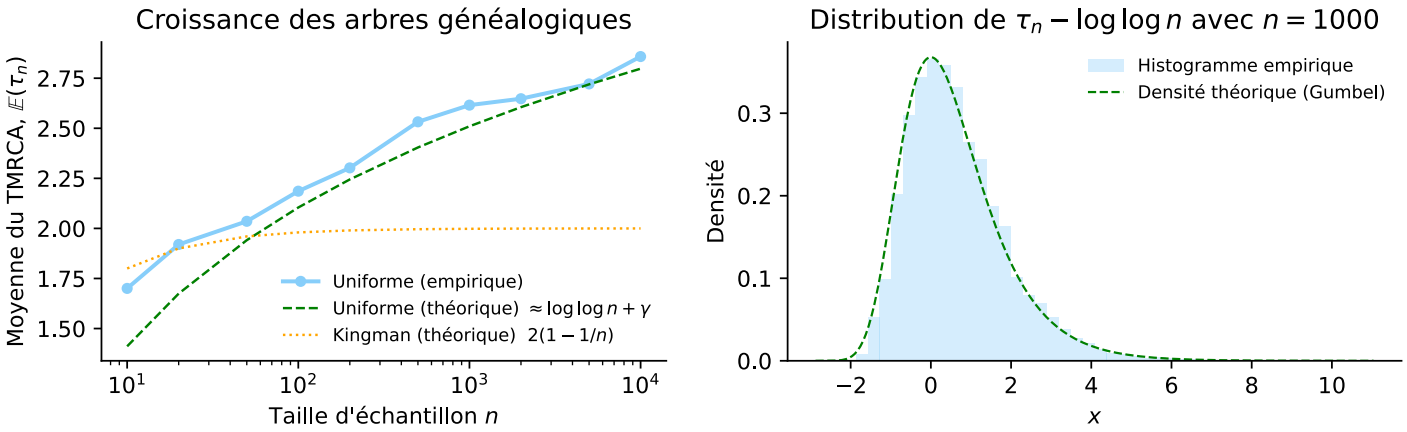

FIGURE 4 – (a) *Gauche* : sous le modèle de Bolthausen-Sznitman (Λ uniforme) nous déterminons les moyennes empiriques de $\mathbb{E}(\tau_n)$ (500 répétitions par $n \in \{10, 20, 50, 100, 200, 500, 1e3, 2e3, 5e3, 1e4\}$) comparées à l'approximation théorique $\gamma + \log\log n$ et aussi à Kingman (2). (b) *Droite* : histogramme de $\tau_n - \log\log n$ pour $n = 1000$ (5000 simulations) avec superposition de la densité de Gumbel $x \mapsto e^{-x-e^{-x}}$.

3 Discordances par paires chez la sardine japonaise

Afin de justifer le modèle du Λ-coalescent, comparé à son prédécesseur le coalescent de Kingman, nous présentons une application en biologie des populations. Nous nous appuyons sur [NNY16] qui étudie la génétique des sardines japonaises. Pour chaque paire de sardines, ils ont analysé leur ADN. En représentant leur séquence comme un mot issu de l'alphabet $\{A, T, G, C\}$, nous comptons pour les lignées $i \neq j$ leur discordance, noté $K_{ij} \in \mathbb{N}$, c'est le nombre de symboles différents lorsque l'on compare leur séquence ADN. Nous notons $T_{ij} \geq 0$ le plus petit temps tel que la lignée i et la lignée j ont un ancêtre commun. Sous des hypothèses biologiques naturelles, nous avons que les mutations ne peuvent pas se produire sur le même site et que pour un paramètre de mutation constant $\mu > 0$, pour tout $t > 0$ [Dur08],

$$K_{ij} \mid T_{ij=t} \sim \text{Poisson}(\mu t)$$

Donc,

$$\mathbb{E}(K_{ij}) = \mu \mathbb{E}(T_{ij}) \tag{4}$$

Ainsi pour modéliser l'arbre généalogique de ces sardines nous devons déterminer un Λ adéquat. Une idée est de comparer les lois marginales prédites par chaque Λ-coalescent pour K_{ij} à la distribution empirique observée.

Comme choix de mesure couramment utilisés en biologie est le modèle de Kingman avec expansion. C'est-à-dire que l'intensité des coalescences varie au cours du temps. Ce modèle permet d'expliquer des populations où le nombre de descendants sont modérés par rapport à la taille de la population totale. Nous considérons dans la suite, par simplicité, le cas de deux époques. Soit $c_0, c_1 > 0$ l'intensité des époques

et $\theta > 0$ le moment de rupture. On définit $c : t > 0 \mapsto c_0 1_{t\leq\theta} + c_1 1_{t>\theta}$ l'intensité de coalescence au cours du temps. Puisque nous avons considéré au début du rapport des mesures de probabilités, afin de faire varier l'intensité de coalescence de la mesure δ_0, nous renormalisons le temps de notre processus selon $s : t \mapsto \int_0^t c(u)du$ et donc on s'intéresse au processus $(N_{s(t)})_{t\geq 0}$. En notant T_K le temps de coalescence entre deux lignées sous ce modèle, nous avons,

$$\mathbb{P}(T_K > t) = \exp\Big(-\int_0^t c(u)du\Big) = \exp(-c_0 \min(t,\theta) - c_1 \max(0, t-\theta))$$

Et donc,

$$\mathbb{E}(T_K) = \int_0^\infty \mathbb{P}(T_K > t)dt = \frac{1}{c_0}(1 - e^{-c_0\theta}) + \frac{e^{-c_0\theta}}{c_1} \tag{5}$$

Ensuite dans la précédente section, sous-section 2.2, nous avons introduit la loi Beta$(2-\alpha, \alpha)$. En partant des mêmes hypothèses que le théorème 1 et supposons que notre population a une reproduction à queue lourde, c'est-à-dire qu'un individu peut être à l'origine d'une partie non négligeable de la population. Alors d'après [Sch04], le processus converge vers un Λ_0^α-coalescent dont la mesure Λ_0^α est la loi Beta$(2-\alpha, \alpha)$ avec $1 < \alpha < 2$. Nous convergeons vers une mesure de probabilité donc en notant T_K^α le temps de coalescence entre deux lignées sous ce modèle, $T_K^\alpha \sim \text{Exp}(r_{2,2}) = \text{Exp}(1)$ donc,

$$\mathbb{E}(T_K^\alpha) = 1 \tag{6}$$

A présent si l'on souhaite utiliser un Λ_0^α-coalescent qui possède une certaine moyenne de discordance entre les lignées, $d_* := \mathbb{E}(K_{ij})$, nous pouvons relier les paramètres des deux modèles par (4), (5) et (6).

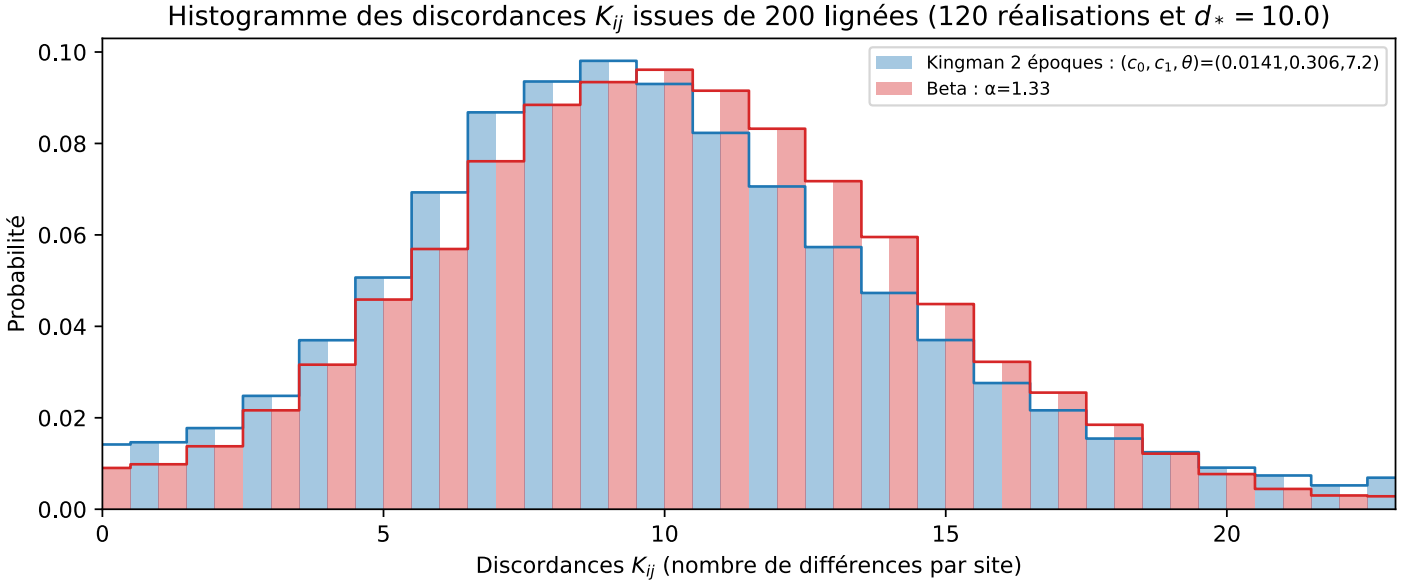

FIGURE 5 – Histogrammes des discordances par paires K_{ij} obtenues à partir de $n = 200$ lignées simulées sous un Λ_0^α-coalescent (en rouge, $\alpha = 1.33$) et sous un coalescent de Kingman à deux époques (en bleu, $(c_0, c_1, \theta) = (0.0141, 0.306, 7.2)$). Ces paramètres ont été choisis pour minimiser la distance en variation totale entre l'histogramme considéré et celui du Λ_0^α-coalescent. Dans les deux modèles, le taux de mutation μ est choisi de sorte que la moyenne des discordances soit fixée à $d_* = 10$. Les histogrammes sont construits à partir de 120 réalisations indépendantes et montrent que les deux modèles produisent des distributions de discordances pratiquement indiscernables.

Dans la figure 5, nous avons deux modèles biologiques plausibles pouvant expliquer la discordance d'une population. Ainsi connaître $(K_{ij})_{i\neq j}$ ne permet pas d'identifier la mesure Λ.

Toutefois, si on fixe le taux de mutation μ pour les deux modèles alors à partir de données, nous pouvons déterminer les paramètres de nos modèles et regarder leur performance. Pour l'expérience numérique nous

prenons $\alpha = 1.3$ comme dans [NNY16] qui a été choisi comme le paramètre maximisant la vraisemblance. Pour le modèle de Kingman avec expansion, ils considèrent un modèle à deux époques et estiment les paramètres démographiques (u_0, u_1), ce sont les tailles effectives actuelles et ancestrales en unités de mutation. L'ajustement par maximum de vraisemblance basé sur la distribution des différences par paires donne $u_0 = 1$, $u_1 = 1.25$ et $\theta = 0.45$ [NNY16]. Dans notre paramétrisation en termes de taux de coalescence, cela correspond à

$$(c_0, c_1, \theta) = \left(\frac{1}{u_0}, \frac{1}{u_1}, \theta\right) = (1, 0.8, 0.45)$$

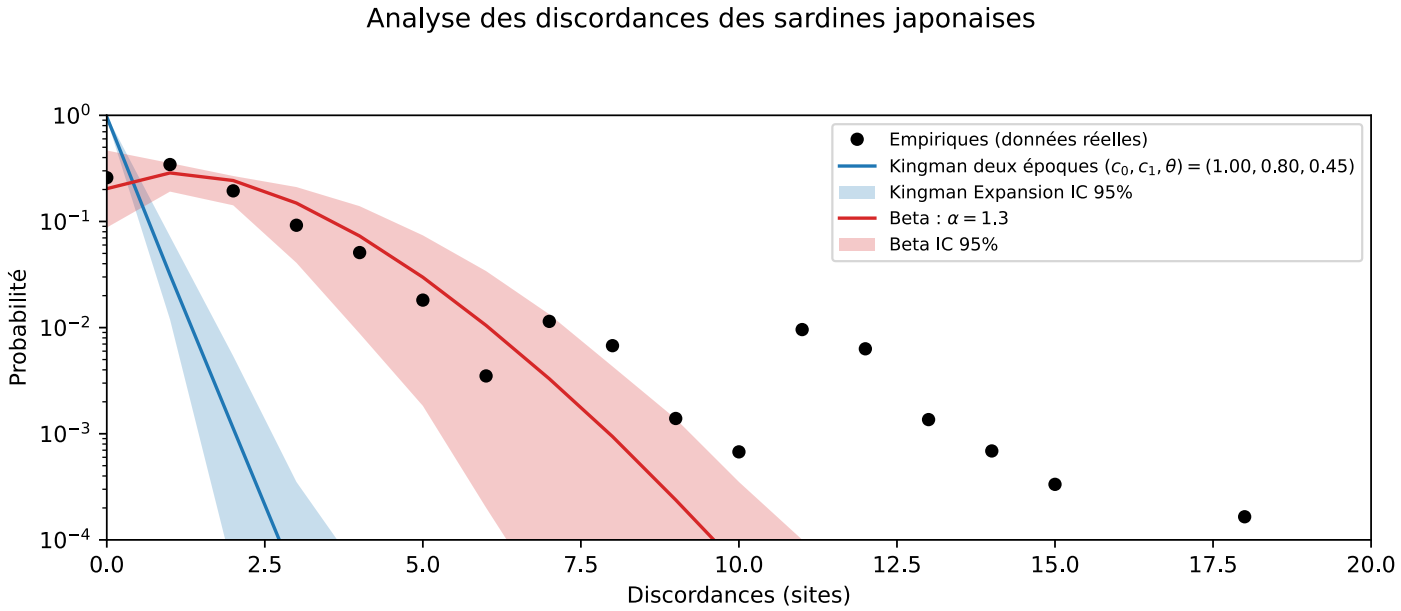

FIGURE 6 – Distribution empirique des discordances par paires chez la sardine japonaise (points noirs) comparée aux prédictions des deux modèles considérés. La courbe rouge correspond au Λ_0^α-coalescent de loi Beta$(2-\alpha, \alpha)$ avec $\alpha = 1.3$, et un intervalle de confiance à 95% en rose obtenu par Monte Carlo (300 simulations). Le taux de mutation μ est calibré de sorte que la moyenne théorique des discordances sous ce modèle coïncide avec la moyenne empirique. La courbe bleue correspond au coalescent de Kingman à deux époques, avec $(c_0, c_1, \theta) = (1.00, 0.80, 0.45)$. La bande bleue indique l'intervalle de confiance à 95%.

On remarque alors que le modèle de Kingman, même avec expansion, ne permet pas d'expliquer la dynamique de ces sardines. Nous avons besoin d'utiliser un modèle permettant la coalescence de multiples lignées à la fois, la mesure Λ_0^α le permet d'une bonne manière et naturellement.

4 Conclusion

Dans ce rapport, nous avons étudié le Λ-coalescent, en particulier le comportement du TMRCA pour différentes mesures Λ et une application aux discordances par paires chez la sardine japonaise. Nos simulations montrent que certains Λ-coalescents à coalescences multiples rendent mieux compte des données que le modèle de Kingman avec expansion.

Nous aurions aimé étudier plus en profondeur pour n fixé et $n \to \infty$ des questions sur le TMRCA et la vitesse de convergence, un aspect qu'on a du délaisser par manque de temps et de place. De plus une autre piste intéressante et d'utiliser un Ξ-coalescent, un modèle plus général permettant la coalescence simultanée.

Références

[Dur08] Richard Durrett. *Probability Models for DNA Sequence Evolution*. Probability and Its Applications. Springer, New York, 2 edition, 2008.

[Fis30] Ronald A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930.

[GM05] Christina Goldschmidt and Jeremy B. Martin. Random recursive trees and the bolthausen–sznitman coalescent. *Electronic Journal of Probability*, 10 :718–745, 2005.

[Kin82] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3) :235–248, 1982.

[KLLS17] Tobias Kluge, Koenigs Leckey, Wolfgang Löhr, and Jason Schweinsberg. Exchangeable coalescents, ultrametrics, and trees. 2017.

[NEM94] Paul Nevai, Tamás Erdélyi, and Alphonse P Magnus. Generalized jacobi weights, christoffel functions, and jacobi polynomials. *SIAM Journal on Mathematical Analysis*, 25(2) :602–614, 1994.

[NNY16] Hiro-Sato Niwa, Kazuya Nashida, and Takashi Yanagimoto. Reproductive skew in japanese sardine inferred from dna sequences. *ICES Journal of Marine Science*, 73(9) :2181–2189, 2016.

[Pit99] Jim Pitman. Coalescents with multiple collisions. *The Annals of Probability*, 27(4) :1870–1902, 1999.

[Sag99] Serik Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36(4) :1116–1125, 1999.

[Sch04] Jason Schweinsberg. Coalescent processes obtained from Cannings models of population reproduction. *Electronic Journal of Probability*, 9 :1–24, 2004. Preprint 2003 ; rigorous derivation of Λ-coalescents from Cannings models.

[Sze39] G. Szegő. *Orthogonal Polynomials*, volume 23 of *Colloquium Publications*. American Mathematical Society, 1939.