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 (Fig. 1). On observe bien que les fusions sont binaires et que le temps total de coalescence oscille autour de la valeur théorique $2(1 - 1/20) = 1.9$.

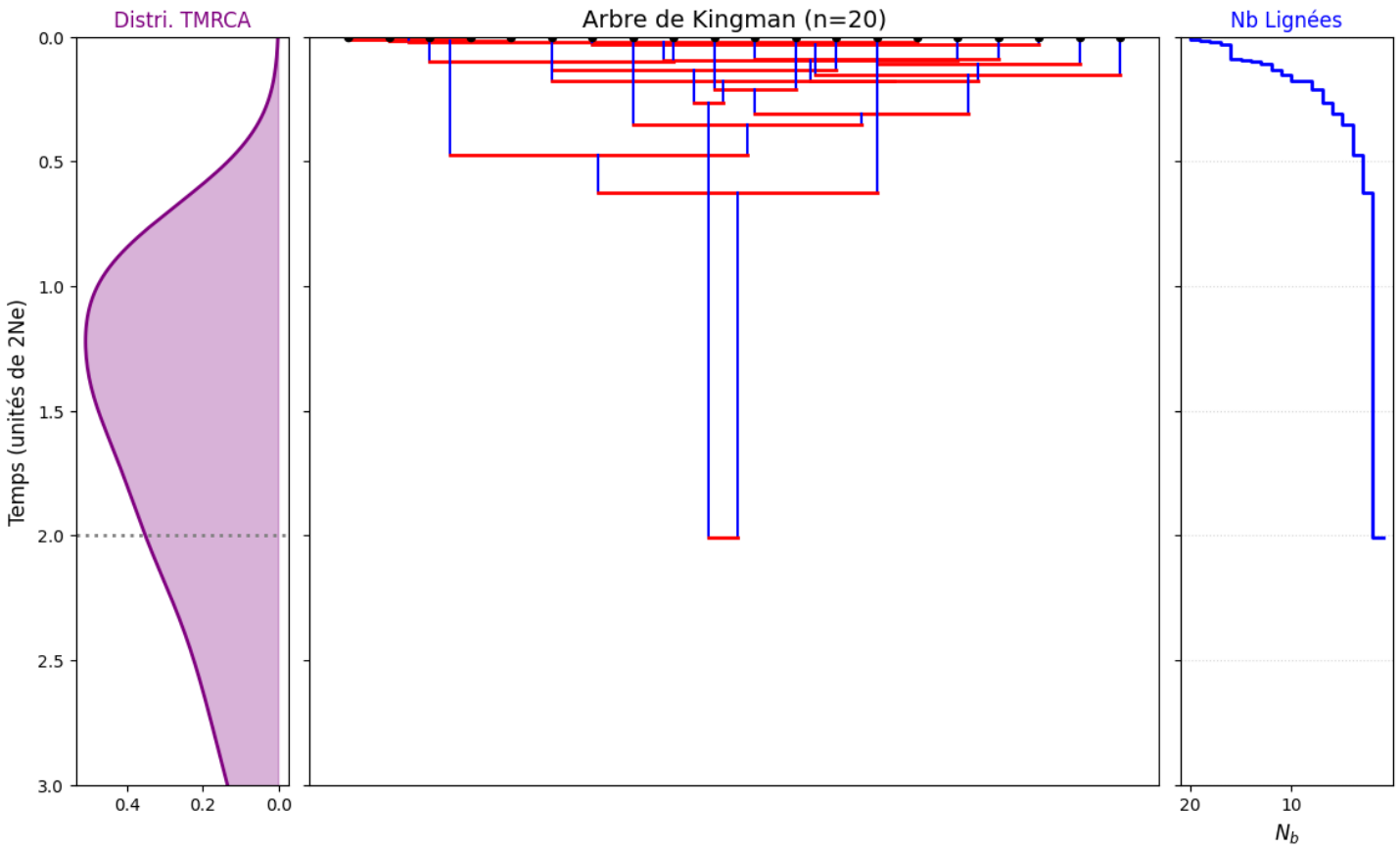

FIGURE 1 – Simulation de Kingman ($n = 20$). **À droite** : Une trajectoire du Kingman-coalescent. **Au centre** : La généalogie réalisée, montrant les fusions par paires pour la trajectoire (à droite). **À gauche** : La densité du TMRCA. La ligne pointillée grise indique le TMRCA de la trajectoire (à droite)(2)et si à la place tu faisait une simulation et générer un histogramme et parler aussi un peu de la comment tu as simuler tous ça .

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 $H(b) \coloneqq \mathbb{E}(\tau_{\Lambda,b})$.

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. □

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 $n \to \infty$.

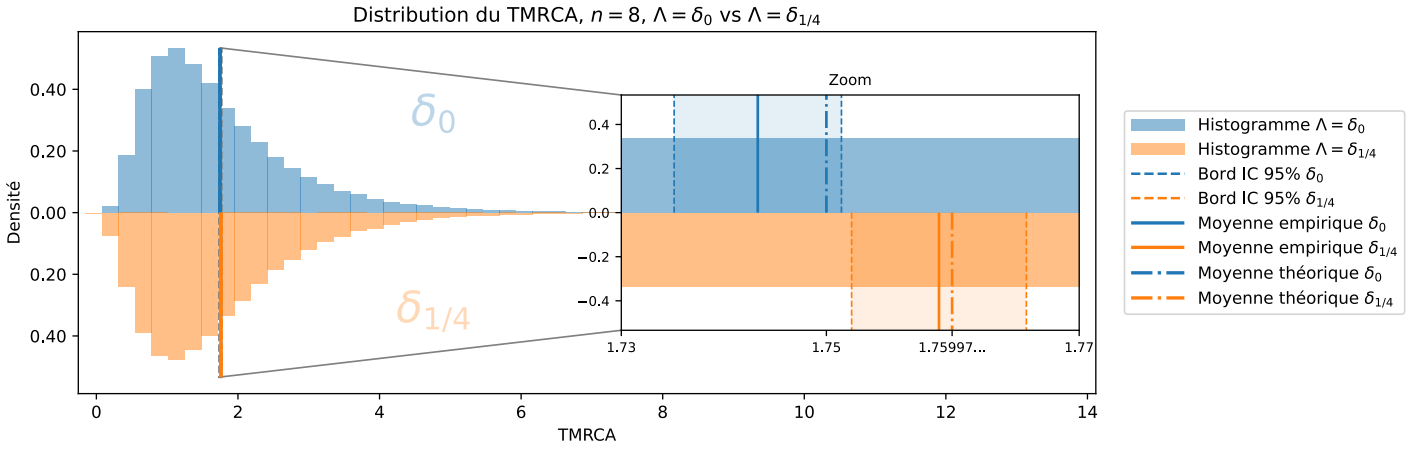

FIGURE 2 – Distribution empirique du TMRCA pour $n = 8$ de $\Lambda = \delta_0$ (Kingman, histogramme vers le haut en bleu) et $\Lambda = \delta_{1/4}$ (histogramme vers le bas en orange). Les bandes verticales pointillées délimitent les intervalles de confiance (IC) à 95% pour la moyenne de τ basés 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 $\mathbb{E}(\tau_{\delta_0,8}) \leq \mathbb{E}(\tau_{\delta_{1/4},8})$.

L'échelle de temps ici est en unités de N générations, avec $N \gg n$ puisque nous considérons un modèle asymptotique.

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 $1 \leq b, i \leq n$, par

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

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 $R = (Q_{i,j})_{2\leq i,j\leq n}$ la matrice devient inversible et nous pouvons exprimer la densité de τ. Posons $p_R(t) = (p_k(t))_{2\leq k\leq n}$ où $p_k : t \geq 0 \mapsto \mathbb{P}(N_t = k \mid N_0 = n)$. D'après la relation de Chapman-Kolmogorov, p_R vérifie pour tout $t \geq 0$

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

Définissons la fonction de survie, $S : t \mapsto \mathbb{P}(\tau_n > t) = \mathbb{P}(N_t \neq 1 \mid N_0 = n) = \sum_{k=2}^n p_k(t) = p_R(t)\cdot\mathbf{1}$. 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 $I_n := \{(b,k), 2 \leq k \leq b \leq n\}$. 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 $(m_r(\Lambda))_{0\leq r\leq n-2}$, 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 $\Lambda_0^\alpha := \mathrm{Beta}(2-\alpha,\alpha)$ avec $\alpha \in]1,2[$, de densité*

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

Soit $(J_n)_{n\geq 0}$ les polynômes de Jacobi. Pour tout $n \geq 0$, $J_n = \sum_{k=0}^{n} \binom{n+\alpha-1}{n-k+1}\binom{n-\alpha+1}{k} x^{n-k}(x-1)^k$ 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

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

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 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 $\beta > -1, \gamma > -1$ (classiquement les polynômes de Jacobi sont définis sur $[-1,1]$ mais par un changement de variable on se ramène sur $[0,1]$).

Le poids w_α implique que $\beta = \alpha - 1, \gamma = 1 - \alpha$ et la condition sur β et γ impose que $\alpha \in]0,2[$. Pour $0 < \alpha < 1$, le Beta$(2-\alpha, \alpha)$ coalescent admet une fraction positive des individus qui reste sous forme de singletons à tout temps t>0 [BBS08] ce qui a pour conséquence de rendre la profondeur des arbres généalogiques explosive quand n tend vers l'infini. Par conséquent, nous gardons $\alpha \in]1,2[$.

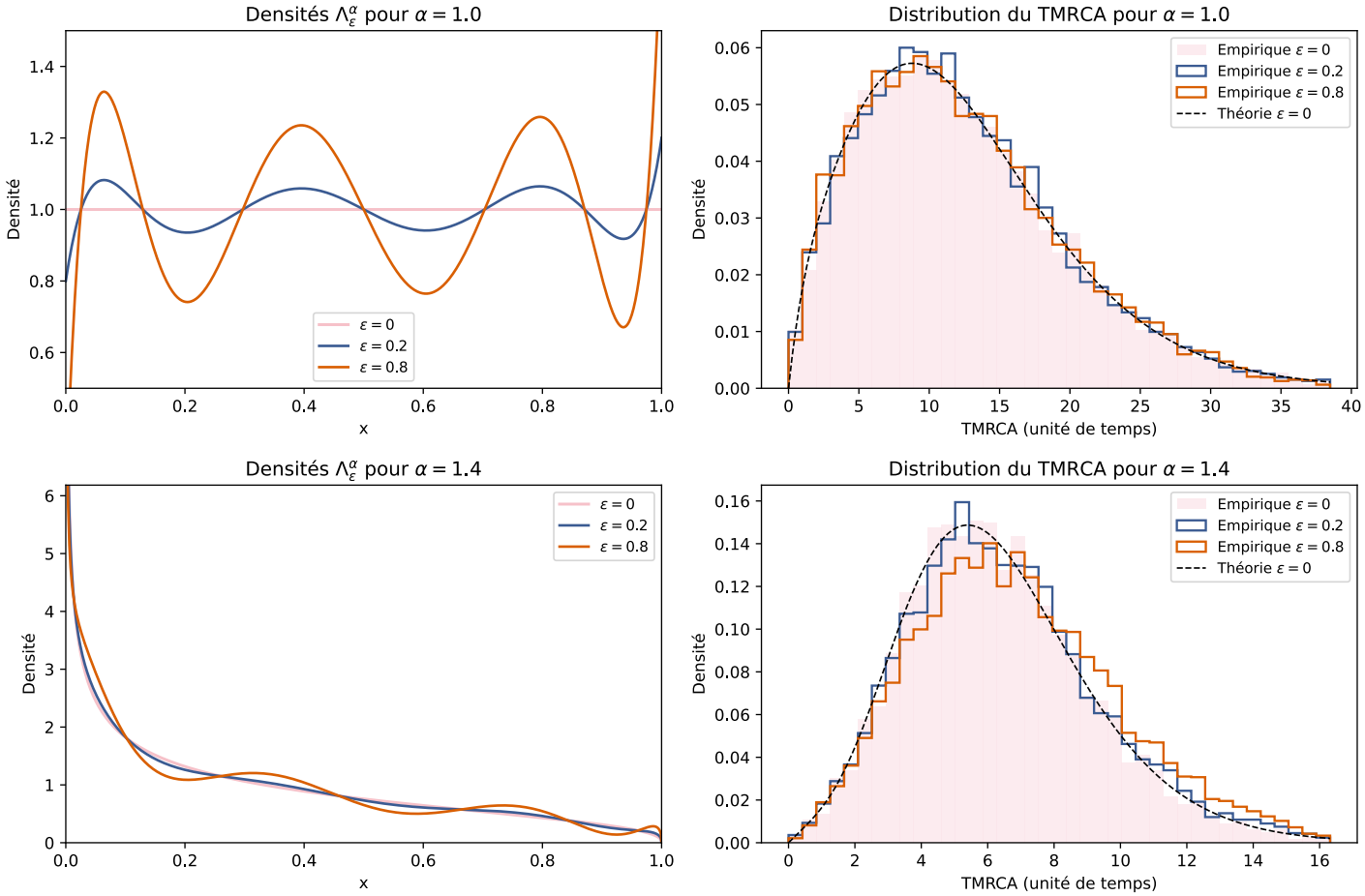

FIGURE 3 – (a) *Gauche* :TODO comparaison des densités remarquablement différentes sur $[0,1]$, $f_\varepsilon(x) = \Big(1 + \varepsilon J_{n-1}(x)\Big)\, w_\alpha(x)$, pour $n = 8$: cas $\mathrm{Beta}(2-\alpha, \alpha)$ (avec $\alpha \in \{1, 1.4\}$)$\varepsilon = 0$ et perturbation de Jacobi $\varepsilon \in \{0.2, 0.8\}$ (b) *Droite* : distribution empirique du TMRCA, issues de n lignées, à partir de 8000 simulations, avec superposition de la densité théorique donnée par (3).

Dire que beta a été simulé par trapeze ? montecarlo ? faire quelques phrase dessus

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 $\alpha = 1.4$, nous avons pour $n = 8$ que $\varepsilon_n = 0.392062$. On remarque alors que la distribution empirique vérifie bien la théorie pour $\varepsilon = 0.2 < \varepsilon_n$ et que pour $\varepsilon = 0.8 > \varepsilon_n$ nous ne convergeons déjà plus en loi.

Dans le cas $\alpha = 1$, graphique du haut, nous obtenons la mesure uniforme et ε_n ne depend plus de n. Cela nous permet d'avoir des densités drastiquement différentes pour une étude initiale de $n \gg 1$ lignées, comparées à une densité avec $\alpha \neq 1$. Le choix de ε est tout de même borné par 1 afin que notre densité soit 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,*

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

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.

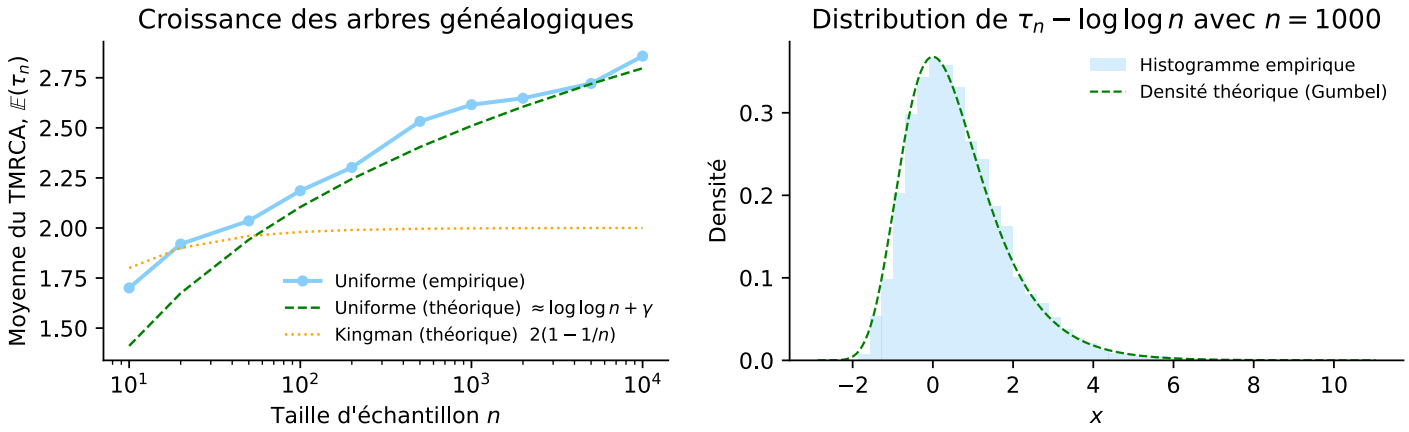

FIGURE 4 – (a) *Gauche* : sous le modèle de Bolthausen-Sznitman (Λ uniforme) nous déterminons les moyennes empiriques de $\mathbb{E}(\tau_n)$ (500 répétitions par $n \in \{10, 20, 50, 100, 200, 500, 1e3, 2e3, 5e3, 1e4\}$) comparées à l'approximation théorique $\gamma + \log\log n$ et aussi à Kingman (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 Application : Le Paradoxe de la Sardine Japonaise

Nous avons précédemment introduit la loi Beta$(2-\alpha, \alpha)$. Celle-ci est cruciale en biologie des populations, car le modèle de Kingman est insuffisant pour décrire certaines espèces.

L'exemple de la sardine japonaise, analysé par Niwa et al. [Niw16], illustre cette limite. Cette étude constitue une démonstration empirique de la nécessité du Λ-coalescent pour les espèces à stratégiedite de *sweepstakes*, dans lesquelles seuls quelques individus parmi des millions réussissent à produire une descendance massive, tandis que la majorité n'en laisse aucune.

Là j'ai du mal car x, C ne sont pas défini. Pourquoi une queue lourde ça le modélise ? car si on veut que la sardine fasse beaucoup de bébés et d'autres de maniere négligeable il faudrait la masse vers 0 et une partie vers 1 mais en soit rien n'oblige à une "queue lourde", peut-être il (me) manque un argument. Mathématiquement, ce comportement est modélisé par une distribution à queue lourde du nombre de descendants : l'article pose explicitement que pourquoi alpha n'a pas le droit d'etre en dessous de 1 ? peut-être cela viendra avec l'explication de la loi beta Vérifier du x->0 pour le Theta

$$\mathbb{P}(X \geq x) = \Theta_{x\to 0}(x^{-\alpha}), \qquad 1 < \alpha < 2,$$

où X désigne le nombre d'enfants produits par un individu. Depuis le début on parle de lignées et là on parle d'enfants. Donc il faut expliquer (en 4 mots, 1 phrases) comment notre modèle de lignées permet de passer à cette interprétation Dans cette plage de valeurs Laquelle ?, la loi possède une variance infinie var(X) ? si oui, à calculer., ce qui formalise le caractère extrême des inégalités reproductives et traduit rigoureusement le phénomène de *sweepstakes*.

3.1 Le Conflit

Chez les animaux à haute fécondité, l'analyse génétique révèle une contradiction si l'on reste dans le paradigme binaire classique (Wright-Fisher ou Kingman) : On observe un excès massif de mutations uniques (présentes sur un seul spécimen).

Sous Kingman, cette observation ne peut être interprétée que comme une expansion démographique explosive et récente (dans cet article, on estime que c'est de l'ordre de 10^3).

Si cette expansion était réelle, la distribution des distances par paires (*mismatch distribution*, définie comme la distribution du nombre de allèles différentes entre deux séquences pour l'ensemble des paires d'individus) devrait former une loi de Poisson. Or, les données montrent une distribution en forme de "L" : un pic en zéro et une queue lourde.

3.2 Résolution par le Beta-Coalescent et Simulation

Le modèle Beta-coalescent (avec $\alpha \approx 1.3$, valeur calculée empiriquement dans l'article) résout ces problèmes.

Les fusions multiples fréquentes créent des groupes d'individus identiques (le pic à 0), tout en laissant survivre des lignées très anciennes (queue lourde), sans recourir à l'hypothèse d'expansion explosive.

Nous avons simulé cette distinction (Fig. 5) en comparant un modèle de Kingman avec expansion et un modèle Beta stationnaire (la réalité).

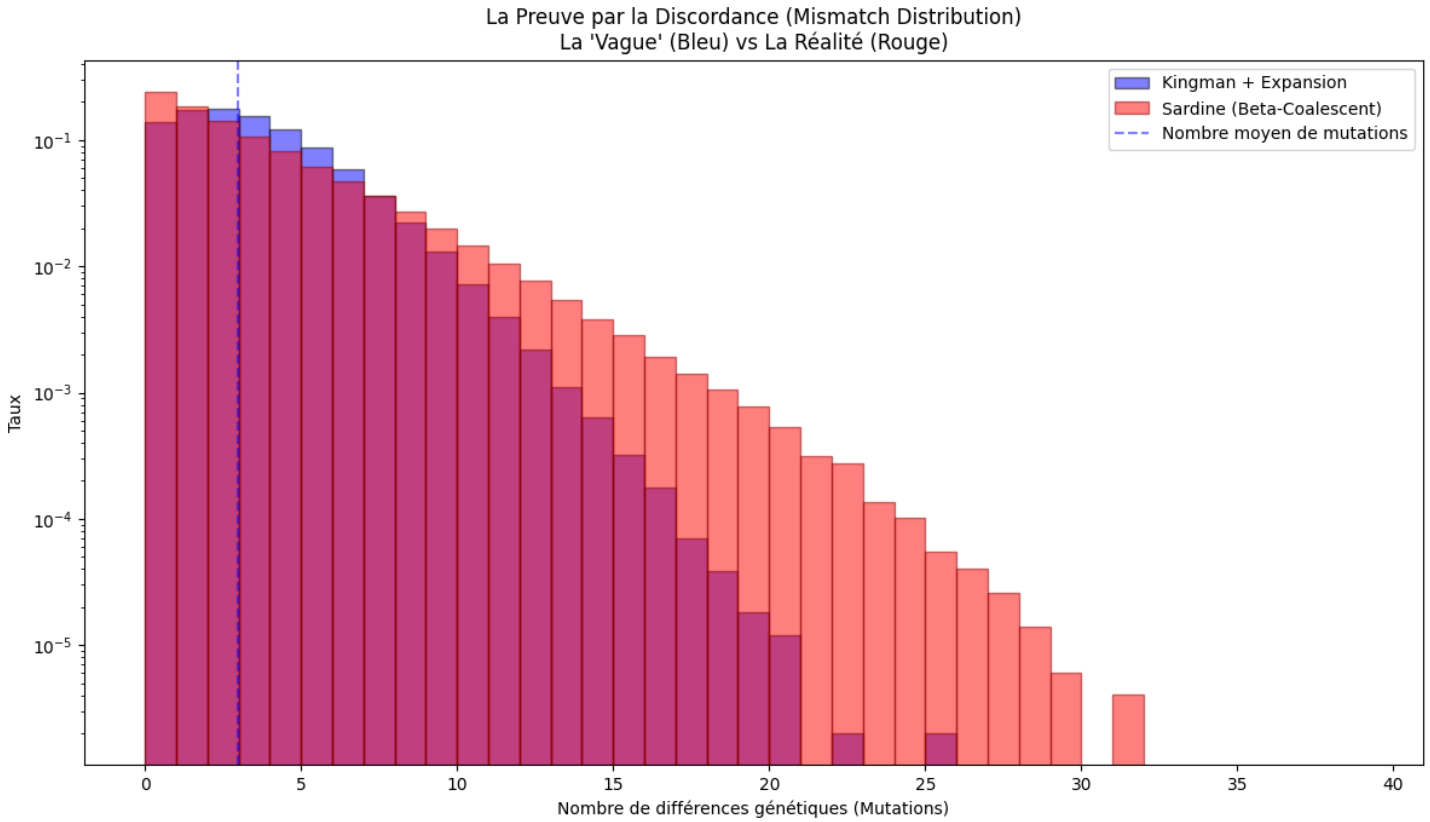

FIGURE 5 – Distribution des discordances ($n = 150$). **En bleu (Kingman + Expansion)** : Une "vague" créée par le phénomène de croissance démographique. **En rouge (Sardine $\alpha = 1.3$)** : Un pic de clones et une queue de lignées persistantes (résultat attendu).

Pour valider cette simulation, nous la confrontons aux résultats empiriques originaux (Fig. 6). On y retrouve exactement la dichotomie observée plus haut.

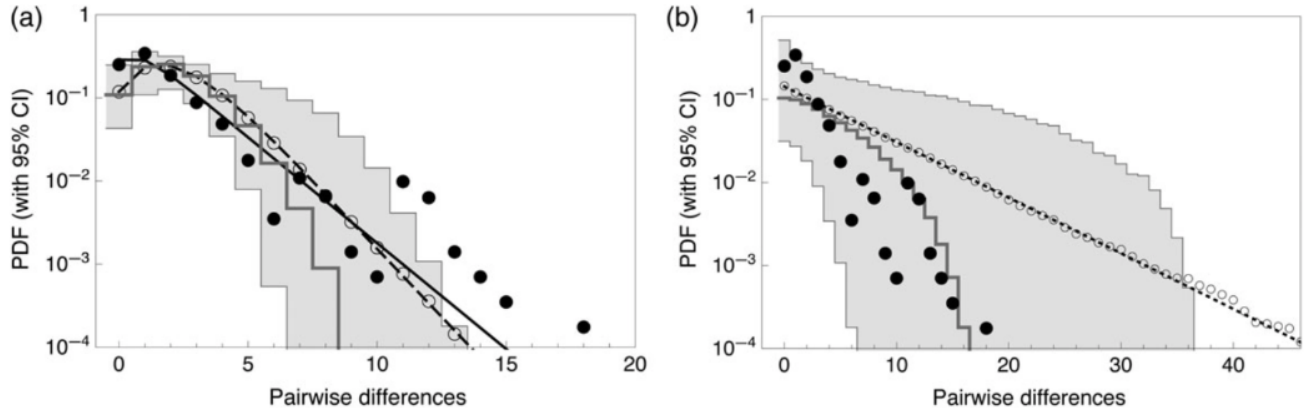

FIGURE 6 – Comparaison des ajustements sur données réelles (Source : Niwa et al., 2016). **(a)** L'approche classique (Kingman), échouant à suivre les points noirs (données empiriques). **(b)** Le modèle Λ-coalescent. (Voir l'article original pour les paramètres complets).

3.3 Conséquence : Millions de sardines, même génome

Ce résultat invalide la relation linéaire entre taille de recensement N et taille efficace N_e. Sous le régime Beta-coalescent, la dérive génétique est régie par une loi de puissance :

$$N_e \propto N^{\alpha-1}$$

Pour la sardine ($\alpha \approx 1.3$), cela implique $N_e \propto N^{0.3}$. Cette relation explique pourquoi des populations marines gigantesques conservent une diversité génétique très faible et instable.

Ce modèle permet non seulement d'expliquer un ancien paradoxe chez les animaux à haute fécondité, mais sert également d'outil pour estimer la taille effective d'une population.

Références

[BBS08] Julien Berestycki, Nathanaël Berestycki, and Jason Schweinsberg. Small-time behavior of beta coalescents. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 44(2), April 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.

[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.

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

[Niw16] H.-S. Niwa, K. Nashida, and T. Yanagimoto. Reproductive skew in Japanese sardine inferred from DNA sequences. *Molecular Ecology*, 25(16) :3831–3843, 2016.