Titre du rapport

Nom de l'auteur

11 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 [Wright-Fisher]. Ensuite, le modèle de Kingman [Kingman] 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 Pitman [1999], Sagitov [1999])**.** *Il existe un processus de Markov, appelé Λ-coalescent, échangeable à collisions multiples simples si et seulement s'il existe une mesure de probabilité 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 est entièrement caractérisée par une mesure de probabilité Λ sur $[0,1]$. Partant de b lignées, le taux total auquel une coalescence se produit, le taux de sortie de l'état b, est donné par

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

Chaque événement de coalescence correspond alors à une transition de b vers $b-k+1$ lignées avec probabilité

$$\forall b \geq k \geq 2, \quad p_{b,k} \coloneqq \frac{\binom{b}{k}\lambda_{b,k}}{\lambda_b}$$

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

1.2 Exemple (Kingman)

Intéressons nous à un cas connu afin de mieux visualiser le modèle. Là je propose donc (1 page grand max) (le but n'est pas de faire une étude de Kingman mais de présenter les différents objets du modèle de manière simple et visuelles)

- poser $\Lambda = \delta_0$,
- l'intuition du modèle : la masse est vers 0 donc pas de grosse fusion. (Inteprétation du modele, du role des termes de l'intégrande)
- Les calculs en 5 lignes maximum ($\lambda_{b,k}$, TMRCA si utile, ...)
- Les notations : n, N_t, $(C_t^i)_{0<i<n}$, TMRCA, T_k

Subplots (kingman)

- (Une réalisation) Arbre + TMRCA
- (Sur plusieurs réalisaiton) Distribution des fusions ((C_t^i))_{0<i<n } (donc uniquement en 2 normalement)
- (Sur plusieurs réalisation) distribution TMRCA
- Autres ?

Le but est de faire un exemple qui montre quasiement tout ce qu'on va étudier et par la suite quantifier pour beaucoup de mesures.

2 Analyse du TMRCA

2.1 Cas extrêmes

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. Notons le TMRCA d'un Λ-coalescent,*

$$\tau := \inf\{t \geq 0, N_t = 1\}$$

Alors, pour toute mesure de probabilité Λ sur $[0,1]$, on a condtionnellement à $\{N_0 = n\}$,

$$1 = \mathbb{E}_{\delta_1}(\tau) \leq \mathbb{E}_\Lambda(\tau)$$

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 $\tau \sim \text{Exp}(1)$ et donc $\mathbb{E}_{\delta_1}(\tau) = 1/1 = 1$.

Soit Λ une mesure de probabilité sur $[0,1]$. Notons $H(b) := \mathbb{E}_\Lambda(\tau | N_0 = b)$. D'après la propriété de Markov et l'absence de mémoire de l'exponentielle,

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

Montrons par récurrence forte l'inégalité, c'est-à-dire $1 \leq H(b)$ pour $b \geq 2$. Pour $b = 2$, $\lambda_{2,2} = 1$ donc $\tau \sim \text{Exp}(1)$ donc $H(2) = 1 \geq 1$. Supposons l'inégalité vraie jusqu'à $b-1$. Donc, en remarquant que $\lambda_{b,b} = \int_0^1 x^{b-2}\,\Lambda(dx) \leq 1$,

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

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 !

Théorème 2. *Il existe $n > 1$ et une mesure de probabilité Λ sur $[0,1]$ telle que, conditionnellement à $\{N_0 = n\}$,*

$$\mathbb{E}_\Lambda(\tau) > \mathbb{E}_{\delta_0}(\tau)$$

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

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

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,

$$\mathbb{E}_{\delta_{1/4}}(\tau) = \frac{1}{\lambda_n} + \sum_{k=2}^{n-1} \frac{\binom{n}{k}\lambda_{n,k}}{\lambda_n}\mathbb{E}_{\delta_{1/4}}(\tau | N_0 = n-k+1) = \frac{19954284839411683}{11337879079537330} > 1.7599662\cdots > 1.75$$

□

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

Là on a une belle expérience numérique à faire ! ! On peut faire un plot de la distribution pour ces 2 mesures et voir qu'en moyenne on a bien delta1/4 devant delta0 ! + intervalle asymtptotique si possible

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

2.2 Effet papillon de Λ

Références

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

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