Titre du rapport

Nom de l'auteur

30 octobre 2025

0.1 Introduction à la théorie de la coalescence.

Nous disposons d'une population haploïde[1] de taille N. On commence à la génération $m = 0$ et nous souhaitons remonter le temps pour étudier la généalogie de $n < N$ individus, généralement on considère $n \ll N$. On appelle la lignée de $i \in [n] := [\![1, n]\!]$ la suite des ancêtres de i. Lorsque deux, ou plus, lignées choisissent le même parent à un instant donné, on dit qu'elles coalescent.

Formellement, on définit pour la génération $m \in \mathbb{N}$, la relation d'équivalence $\sim_m$ pour tout $i, j \in [n]$,

$$i \sim_m j \text{ ssi les lignées de } i \text{ et } j \text{ ont le même ancêtre à la génération m}$$

La classe de i à la génération m, constituant un bloc, est,

$$C_m^{(N)}(i) := \{j \in [n], i \sim_m j\}$$

Puis, la famille des classes distinctes forme une partition de [n]

$$\Pi_m^{(N)} := \{C_m^{(N)}(i), i \in [n]\} \qquad K_m := |\Pi_m^{(N)}|$$

Dans le modèle historique de Wright-Fisher, à chaque génération, chaque individu choisit un parent au hasard dans la génération précédente, de manière indépendante et uniforme. On a ainsi que $(\Pi_m^{(N)})_{m \geq 0}$ est une chaîne de Markov homogène à temps discret sur les partitions de $[n]$.

Notons $(\cdot)_k = \prod_{i=\cdot-k+1}^{\cdot} i$. On a que la probabilité que $l \geq 2$ lignées de k blocs coalescent en une génération est donnée par,

$$\binom{k}{l} \frac{(N)_{k-l+1}}{N^k} = \begin{cases} \binom{k}{2} \frac{1}{N} + \mathcal{O}(\frac{1}{N^2}) & \text{si } l = 2 \\ \mathcal{O}(\frac{1}{N^2}) & \text{si } k \geq l \geq 3 \\ 0 & \text{sinon} \end{cases}$$

On remarque donc que l'évènement d'une coalescence est rare lorsque N est grand. On considère ainsi le processus de Markov $(\Pi^{(N)}(t))_{t \geq 0}$ défini par,

$$\Pi^{(N)} : t \in \mathbb{R}^+ \longmapsto \Pi_{\lfloor tN \rfloor}^{(N)}$$

Notons le semi-groupe associé

$$P_t^N : (x, y) \mapsto \mathbb{P}(\Pi^{(N)}(t) = y | \Pi^{(N)}(0) = x)$$

Considérons x une partition de $[n]$ avec k blocs, y une partition de $[n]$, avec $k-1$ blocs, obtenue en fusionnant deux blocs de x, et y^+ une partition de $[n]$ en fusionnant plus de deux blocs de x. Puisque $t \mapsto P_t^N$ est en escalier, le générateur infinitésimal est donné par,

$$q^{(N)} : (x, z) \longmapsto N(P_{1/N}^N(x, z) - \delta_{x,z}) = \begin{cases} N\left(\frac{1}{N} + \mathcal{O}(\frac{1}{N^2}) - 0\right) = 1 + \mathcal{O}(\frac{1}{N}) & \text{si } z = y \\ N\left(\mathcal{O}(\frac{1}{N^2}) - 0\right) = \mathcal{O}(\frac{1}{N}) & \text{si } z = y^+ \\ N\left(1 - \binom{k}{2}\frac{1}{N} + \mathcal{O}(\frac{1}{N^2}) - 1\right) = -\binom{k}{2} + \mathcal{O}(\frac{1}{N}) & \text{si } z = x \end{cases}$$

En passant à la limite sur N vers l'infini, nous obtenons le coalescent de Kingman. Son générateur est donné par,

$$q : (x, z) \longmapsto \lim_{N \to \infty} q^{(N)}(x, z) = \begin{cases} 1 & \text{si } z = y \\ -\binom{k}{2} & \text{si } z = x \\ 0 & \text{sinon} \end{cases}$$

Ce modèle considère uniquement le fait que deux lignées peuvent coalescer à la fois. Nous pourrions envisager des généralisations où plusieurs lignées coalescent simultanément. [INTRODUIRE NATURELLEMENT LE LAMBA-COALESCENT]

TODO :

1. Chaque individu porte une seule copie du génome, modélisée par un unique parent.

— Image vulgarisatrice de ce qu'on fait (j'ai une idée donc à voir si c'est utile)
— Détails des calculs d'Axcel pour différentes mesure : nouvelle section + subsections
— Enchainer avec une succession de beau plot
— lancer une problématique (à en discuter voir où on veut aller : trouver un parametre par simulation (temps d'un unique ancetre, pb de coalescence multiple, ...), estimer la distance entre deux modèles (KS, Wasserstein, ...), usage détourner de la coalescence pour faire quelquechose d'atypiques)