

Coalescent de Kingman

Gabriel Lang

Université Paris-Saclay, AgroParisTech

MAEE, master BEE

On considère une génération à k individus. Construire leur généalogie consiste à déterminer toutes les coalescence les concernant dans le passé.

- Quelle est la probabilité d'avoir une coalescence à la génération précédente ?
- Quelle loi de probabilité suit le temps de coalescence $G \in \mathbb{N}^*$?

B épreuve de Bernoulli de probabilité de succès p . On répète le tirage de façon indépendante jusqu'au premier succès à la date G .

Définition de la loi géométrique de probabilité de succès p

$$\mathbb{P}(G = k) = p(1 - p)^{k-1}, \mathbb{P}(G \geq k) = (1 - p)^k \text{ et } E(G) = 1/p$$

Propriété

G_1 et G_2 indépendantes de loi géométrique de paramètre p_1 et p_2 . Soit $G = \min(G_1, G_2)$. Alors G est une loi géométrique de paramètre $p_1 + p_2 - p_1p_2$. De plus $\mathbb{P}(G = G_1) = p_1/(p_1 + p_2)$ et $\mathbb{P}(G = G_2) = p_2/(p_1 + p_2)$

Temps de coalescence du modèle de Wright Fisher

On suppose que

- L'effectif de la génération $2N$ est grand.
- L'effectif k de l'ensemble d'individus étudié est petit par rapport à $2N$.

Les événements de coalescence sont rares. Le temps pour observer la première coalescence est long.

Temps de première coalescence G exprimé en nombre de générations

Le temps de coalescence G calculé en générations suit une loi géométrique de paramètre $k(k-1)/4N$.

$$\mathbb{P}(G = g) = \frac{k(k-1)}{4N} \left(1 - \frac{k(k-1)}{4N}\right)^{g-1}.$$

Variable T de loi exponentielle de paramètre λ

Pour $t > 0$, $f_T(t) = \lambda e^{-\lambda t}$, $Pr(T \geq t) = e^{-\lambda t}$ et $E(T) = 1/\lambda$.

Propriété

T_1 et T_2 indépendantes de loi exponentielle de paramètre λ_1 et λ_2 . Soit $T = \min(T_1, T_2)$. Alors T est une loi exponentielle de paramètre $\lambda = \lambda_1 + \lambda_2$.

Preuve : $\mathbb{P}(T \geq t) = \mathbb{P}(T_1 \geq t)\mathbb{P}(T_2 \geq t) = e^{-(\lambda_1 + \lambda_2)t}$.

Processus de Markov à temps continu

Exponentielle de matrice

Soit Q une matrice carrée. On définit $e^Q = \sum_{i=0}^{\infty} \frac{Q^i}{i!}$.

Définition du processus de Markov

Soit Q une matrice carrée dont les coefficients non diagonaux sont positifs. On suppose que la somme par ligne est nulle. Alors e^{tQ} est une matrice stochastique pour tout $t \geq 0$. On définit le processus de Markov $(X_t)_{t \in \mathbb{R}}$ par la donnée de L_0 loi de X_0 et de la matrice des lois conditionnelles $\mathcal{L}(X_t|X_s) = e^{(t-s)Q}$. La loi de X_t est $L_t = L_0 e^{tQ}$.

Note : pour $i \in \mathbb{N}$, les X_i forment une chaîne de Markov de matrice de transition $P = e^Q$.

Factorisation du processus de Markov

On peut décrire l'évolution du processus de Markov comme le tirage de deux variables indépendantes :

Temps de saut

Soit $s \in \mathbb{R}$. On suppose que $X_s = i$. Le processus reste dans l'état i pendant un temps T aléatoire exponentiel de paramètre $|q_{i,i}| = \sum_{j \neq i} q_{i,j}$. A l'instant $s + T$, il saute sur un des états $j \neq i$.

Choix du saut

Si le processus est dans l'état i , le saut j est choisi avec probabilité $q_{i,j}/|q_{i,i}|$. La suite des états choisis par les sauts forme une chaîne de Markov indépendante du temps de saut.

Interprétation : au temps s , on lance $k - 1$ variables exponentielles de paramètres $q_{i,j}$. Le saut choisi est celui qui correspond au plus petit temps et ce temps minimum est une variable exponentielle de paramètre $|q_{i,i}|$.

Passage en temps continu du modèle de Wright Fisher pour N infini

On change d'échelle de temps. Si g est le nombre de génération on pose $t = g/2N$. Le temps $t = 1$ correspond donc à $2N$ générations.

Temps de première coalescence T_1 exprimé dans la nouvelle unité de temps

$$\mathbb{P}(T_1 > t) = \mathbb{P}(G > 2Nt) = \left(1 - \frac{k(k-1)}{4N}\right)^{2Nt}$$

Si N tend vers l'infini, $\mathbb{P}(T_1 > t) \rightarrow \exp\left(\frac{-k(k-1)t}{2}\right)$. La loi du temps de première coalescence tend vers une loi exponentielle de paramètre $k(k-1)/2$.

- L'espérance du temps de coalescence est $2/k(k-1)$.
- La coalescence associe une paire d'individus choisie équiprobablement.
- Après la coalescence, la population étudiée à un effectif de $k-1$.

Coalescences successives du modèle de Wright Fisher pour N infini

Le temps de deuxième coalescence T_2 mesuré depuis T_1 correspond au temps de première coalescence pour une population d'effectif $k - 1$.

Temps de coalescence T_i

Le temps T_i de i -ème coalescence est une variable exponentielle de paramètre $(k - i + 1)(k - i)/2$.

$$\mathbb{P}(T_i > t) \rightarrow \exp\left(\frac{-(k - i + 1)(k - i)}{2}\right)$$

- L'espérance du temps de i -ème coalescence est $2/(k - i + 1)(k - i)$.
- Cette espérance augmente avec i .
- La coalescence touche deux lignages choisis de façon équiprobable.

Coalescent de Kingman

Le processus de coalescence du modèle de Wright Fisher pour N infini est un processus de Markov à temps continu appelé Coalescent de Kingman.

Etats du coalescent de Kingman

On considère un ensemble E de n individus. Le coalescent de Kingman décrit la façon dont les individus trouvent un lien de parenté dans le passé.

- A l'étape 0 (dernière génération), les individus sont isolés.
- A l'étape 1, le coalescent rassemble les individus ayant le même père.
- A l'étape 2 le coalescent rassemble les individus ayant le même grand-père.
- A l'étape $n-1$, tous les individus forment un seul ensemble.

Définition de l'ensemble des états

Une partition de E est un ensemble de sous-ensembles de E deux à deux disjoints et d'union E . On appelle $\mathcal{Q}$ l'ensemble de toutes les partitions de E .

Pour $q \in \mathcal{Q}$, on définit k_q le nombre d'ensemble formant la partition q .

Definition du coalescent de Kingman

Définition du processus

On définit un processus de Markov K à temps continu par ses temps d'attente et ses sauts :

- L'espace d'état de K est Q .
- Au temps 0 l'état est fixé à q_0 où q_0 est formé de n singletons.
- A chaque saut l'état passe de q à q' , où q' est une des partitions obtenues par réunion de deux sous-ensembles faisant partie de la partition q . Le nombre de ces partitions possibles est $k_q(k_q - 1)/2$. Le saut choisit de manière équiprobable.
- Le temps d'attente du saut depuis une partition q est une variable exponentielle de paramètre $k_q(k_q - 1)/2$

Le générateur infinitésimal de ce processus est simple : les éléments non diagonaux valent 1 si le saut est possible et 0 sinon.

Propriétés du coalescent de Kingman

On considère la suite S des états visités par la chaîne et on les numérote par ordre temporel. S est une chaîne de Markov non stationnaire, car la chaîne part de la partition p_0 et termine en $n - 1$ étapes sur la partition formée de l'ensemble E seul.

Loi de la chaîne des états

Soit S_i l'état de la chaîne après le i -ème saut. Soit q une partition à $n - i$ ensembles d'effectifs respectifs a_j . Alors

$$P(S_i = q) = \frac{k!(n - k)!(n - k - 1)!}{n!(n - 1)!} a_1! a_2! \dots a_{n-i}!$$

La démonstration se fait par récurrence à partir de l'état initial.

Processus des effectifs

On considère la variable $R = k_q$ au cours du temps. R définit un processus de Markov :

- L'espace d'état de R est $\{1, \dots, n\}$.
- Au temps 0, $R_0 = n$.
- A chaque saut, R décroît d'une unité (processus de mort pur). Il n'y a qu'un saut possible.
- Le temps d'attente du saut depuis R_i est une variable exponentielle de paramètre $R_i(R_i - 1)/2$

Dans le processus de Kingman, la chaîne des sauts et le processus des temps d'attente sont deux processus de Markov indépendants.

Représentation du coalescent de Kingman

Afin de représenter à la fois le processus des temps d'attente et les coalescences d'ensemble, on adopte une représentation graphique de type arbre ou peigne, pour retrouver un arbre de généalogie. Le temps est représenté verticalement les partitions horizontalement

- En partant du bas, on dessine n branches parallèles verticales de longueur T_1 représentant la généalogie des n individus de départ
- Au bout de ces segments correspondant l'instant de coalescence, on joint les deux branches coalescentes par un segment horizontal; ce segment sert de départ à une branche unique. Pour plus de clarté, on réordonne les branches pour que les deux branches coalescentes soient voisines.
- On repart avec $n - 1$ branches verticales de longueur T_2 et ainsi de suite.
- On finit en joignant le haut des deux segments de longueur T_{n-2} par le dernier segment correspondant à l'ancêtre commun des n individus.

Hauteur du coalescent de Kingman

On étudie les caractéristiques géométriques du coalescent.

Hauteur de l'arbre

La hauteur H est la somme des temps de coalescence.

$$\mathbb{E}(H) = \sum_{i=1}^{n-1} \mathbb{E}(T_i) = \sum_{i=1}^n \frac{2}{(n-i)(n-i+1)} = 2 \sum_{i=1}^{n-1} \left(\frac{1}{n-i} - \frac{1}{n-i+1} \right) = 2 \left(1 - \frac{1}{n} \right).$$

$$\text{Var}(H) = \sum_{i=1}^{n-1} \text{Var}(T_i) = \sum_{i=1}^{n-1} \frac{4}{(n-i)^2(n-i+1)^2} \rightarrow 1.15.$$

La hauteur de l'arbre ne varie pas beaucoup avec n . Les branches du bas sont très courtes, car le temps de coalescence est petit quand l'effectif est important. Les branches du haut sont les plus longues et décident de la hauteur totale et de sa variabilité.

Longueur totale du coalescent de Kingman

La hauteur L est la somme des temps de coalescence multipliés par le nombre de branches actives.

Longueur totale des branches de l'arbre

Pour $i = 1, \dots, n-1$, la longueur totale des branches de longueur T_i est $L_i = (n-i+1)T_i$.

- L_i est une variable exponentielle de paramètre $(n-i)/2$.
- La loi de dépassement de L est $\mathbb{P}(L < t) = \left(1 - e^{-\frac{t}{2}}\right)^{n-1}$.
- $\mathbb{E}(L) = \sum_{i=1}^{n-1} \mathbb{E}(L_i) = \sum_{i=1}^{n-1} \frac{2}{n-i} \approx 2 \log(n)$.
- $\text{Var}(L) = \sum_{i=1}^{n-1} \text{Var}(L_i) = 4 \sum_{i=1}^{n-1} \frac{1}{(n-i)^2} \rightarrow \frac{4\pi^2}{6}$.

La longueur de l'arbre est importante car c'est le paramètre qui joue sur le nombre moyen de mutations apparaissant dans la généalogie.

Taux de mutation dans le modèle de Wright Fisher

On considère une population actuelle de gènes inhomogène. Le modèle précédent de Wright-Fisher n'est pas adapté car la population finale de gènes est issue d'un individu unique et est de ce fait homogène. On modifie le modèle en introduisant la possibilité de mutation des gènes au cours du temps.

On considère qu'un individu a une petite probabilité u de muter indépendamment de toute généalogie (mutation neutre).

- En partant de la dernière génération, et en remontant la généalogie, à chaque génération on tire une variable de Bernoulli de probabilité u sur chaque branche.
- La génération correspondant à la première mutation en remontant une branche suit une loi géométrique de paramètre u .

Taux de mutation dans le coalescent de Kingman

Si $u \gg 1/N$, les mutations sont trop rapides: tous les individus sont différents.

Si $u \ll 1/N$, les mutations sont négligeables et les individus sont identiques.

Taux de mutation

Par convention on pose $u = \theta/4N$. On effectue le changement de temps $t = g/2N$

- Le temps d'attente de mutation d'une branche est une variable exponentielle de paramètre $\theta/2$
- Le temps d'attente de la première mutation d'une des n branches est une variable exponentielle de paramètre $n\theta/2$
- La branche mutée est choisie de façon équiprobable.
- Le temps d'attente du premier événement de coalescence ou de mutation est une variable exponentielle de paramètre $n(n-1)/2 + n\theta/2$.

Les mutations sont indépendantes des coalescences ; on peut donc tirer entièrement la géométrie du coalescent puis tirer le processus de mutation en remontant les branches.

Du fait du taux de mutation choisi, les k individus de la sous population observée sont différents ; l'observation de leur différence est en contradiction avec certains arbres généalogiques. On choisit donc de modifier le coalescent en fonction de l'information disponible chez les individus observés ; les coalescences et mutations ne sont plus indépendantes puisque leur ordre doit être cohérent avec la population finale.

Généalogie d'une population hétérogène de gènes

Selon la qualité des observations, on peut choisir **deux descriptions** de l'effet des mutations:

Modèle d'ensemble d'allèles infini

On suppose qu'on observe l'allèle du gène globalement.

- Chaque mutation remplace le gène courant de la branche par un gène nouveau.
- Les anciens gènes ne réapparaissent jamais (stock infini d'allèles)

Si deux mutations ont lieu sur une même branche, on ne connaît que la dernière.

Modèle d'ensemble de sites infini

On suppose qu'on peut séquencer les allèles. On observe les sites le long de l'allèle.

- Chaque mutation modifie un site sur le gène courant de la branche.
- Les sites déjà modifiés une fois ne sont plus modifiés (stock infini de sites)
- Il n'y a qu'une modification possible par site.

Toutes les mutations de la généalogie sont observables dans la population présente.

Modèle d'ensemble d'allèles infinis

Dans ce modèle, on peut considérer que les allèles sont d'une couleur particulière, et que chaque mutation change en une nouvelle couleur. Les couleurs disparues ne reviennent jamais. Le modèle de référence est un modèle de coalescent remontant avec mort des branches car on cesse d'observer les branches dès qu'elles portent une mutation. On peut pas remonter la généalogie au delà d'une mutation.

Événements possibles

On assemble les n individus par couleur. En remontant le temps deux événements sont possibles :

- Coalescence dans une couleur : l'ensemble correspondant contient au moins deux éléments et perd un élément.
- Mutation : le dernier individu d'une couleur disparaît. La branche correspondante disparaît.

Modèle descendant : Urne de Hoppe

L'urne de Hoppe est un modèle pour générer un coalescent descendant coloré. A l'étape 1 on place une boule de couleur dans une urne avec une boule noire. Par la suite, quand on tire une boule de couleur, on ajoute une boule de même couleur, quand on tire la boule noire on ajoute une boule d'une nouvelle couleur; si le nombre de boules dans l'urne est $k + 1$, la chance de tirer une boule de couleur est $1/(k + \theta)$ la chance de tirer la noire est $\theta/(k + \theta)$.

Coalescent coloré

- A l'étape k , les boules colorées représentent les couleurs des k branches présentes.
- L'apparition d'une nouvelle couleur correspond à une mutation et au point de départ d'une famille qui n'est pas reliée à la généalogie courante
- L'apparition d'une nouvelle boule dans une couleur correspond à une coalescence dans la branche de cette couleur.

Ce modèle permet de définir les probabilités descendantes des mutations et coalescences et de prouver la formule d'Ewens qui donne la probabilité des répartitions d'effectifs par couleur. C'est la seule information pertinente, les couleurs ayant toutes la même signification.

Formule d'Ewens

On considère une population actuelle à n individus. On décrit l'état par le vecteur $(a_1, a_2, \dots, a_k)$ où a_i est le **nombre** de couleurs qui sont répétées i fois. Ce vecteur vérifie $\sum_i i a_i = k$ lorsqu'il y a k branches dans le coalescent. On note $P_k(a)$ la probabilité d'observer l'état a quand on a k branches.

La formule d'Ewens donne la probabilité des effectifs par couleur lorsqu'il y a k branches.

On peut calculer à la main les premiers termes

- $P_1(1) = 1$
- $P_2(2, 0) = \frac{\theta}{1+\theta}$, $P_2(0, 1) = \frac{1}{1+\theta}$
- $P_2(3, 0, 0) = \frac{\theta}{1+\theta} P_2(2, 0)$, $P_3(1, 1, 0) = \frac{2}{2+\theta} P_2(2, 0) + \frac{\theta}{2+\theta} P_2(2, 0)$,
 $P_3(0, 0, 1) = \frac{2}{2+\theta} P_2(0, 1)$.

La formule d'Ewens donne la probabilité des effectifs par couleur lorsqu'il y a k branches.

Probabilité d'Ewens

$$P_k(a) = \frac{k!}{(\theta(\theta+1) \cdots (\theta+k-1))} \prod_j \frac{1}{a_j!} \left(\frac{\theta}{j}\right)^{a_j}$$

Nombre d'allèles

Un nouvel allèle apparaît à chaque tirage de la boule noire. Le nombre d'allèles final K_n correspond à la somme de n variables de Bernoulli indépendantes mais de paramètres différents.

Statistique du nombre d'allèles

La probabilité d'une nouvelle couleur à l'étape i est $\frac{\theta}{\theta+n-i}$.

$$\mathbb{E}(K_n) = \sum_{i=1}^n \frac{\theta}{\theta+n-i} \approx \theta \log(n)$$

$$\text{Var}(K_n) = \sum_{i=1}^n \frac{\theta}{\theta+n-i} \left(1 - \frac{\theta}{\theta+n-i}\right) \approx \theta \log(n)$$

- $K_n/\log(n)$ est un estimateur de θ .
- Cet estimateur est très lent : sa variance est $1/\log(n)$ (au lieu de $1/n$).
- K_n est une statistique exhaustive : elle contient toute l'information que peut procurer

Modèle d'ensemble de sites infinis

Dans ce modèle on observe les sites sur les allèles d'un même gène; les mutations ne peuvent affecter un site qu'une fois ; on ne considère que les sites faisant l'objet d'une mutation sur le gène et on note 1 quand le site est affecté, 0 sinon. Ex : 2 mutations

- Mutations sur 2 branches : l'échantillon final peut contenir $\{00, 01, 10\}$
- Mutations sur la même branche : l'échantillon final peut contenir $\{00, 01, 11\}$ ou $\{00, 10, 11\}$
- L'échantillon ne peut pas contenir les quatre allèles $\{00, 01, 10, 11\}$

L'information dans ce modèle est plus importante :

- On peut mesurer une distance entre deux allèles (nombre de mutations).
- Certaines populations d'allèles sont impossibles.
- Toutes les mutations sont visibles.

Estimateur du taux de mutation : nombre de mutations dans la population finale

Dans la population finale de taille n , on observe S_n sites mutés. Ce nombre correspond au nombre total de mutations apparues dans le coalescent.

Nombre de mutations dans la population de taille n

- Conditionnellement au coalescent S_n est une variable de Poisson de paramètre θL_n
- $\mathbb{E}(S_n) = \theta \sum_{i=2}^n n - 1 \frac{1}{i} \approx \theta \log(n)$.
- $\text{Var}(S_n) = \theta \sum_{i=2}^n n - 1 \frac{1}{n-i} + \theta \sum_{i=2}^n n - 1 \frac{1}{(n-i)^2} \approx \theta \log(n)$.

Cette statistique est peu efficace car de variance trop grande.

Estimateur du taux de mutation : nombre de mutations pour une paire

On considère les paires d'individus dans la population finale. Pour une population à deux individus le nombre S_2 de mutations observables correspond aux différences de site. Ce sont les mutations avant le moment de coalescence quand on remonte le temps.

Nombre de différence pour une paire

$$\mathbb{P}(S_2 = k) = \left(\frac{\theta}{\theta + 1} \right)^k \frac{1}{\theta + 1}$$

$$\mathbb{E}(S_2) = \theta \text{ et } \text{Var}(S_2) = \theta + \theta^2$$

Estimateur du taux de mutation : nombre de mutations par paire

Soit $\{i, j\}$ une paire extraite de la population n ; on note $S_2(i, j)$ le nombre de différences observées entre i et j

Estimateur de Najima

On considère l'ensemble des paires et on définit l'estimateur empirique de Najima par :

$$\hat{\pi} = \frac{2}{n(n-1)} \sum_{i \neq j} S_2(i, j)$$

$$\mathbb{E}(\hat{\pi}) = \theta.$$

$$\text{Var}(\hat{\pi}) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$$

La variance de $\hat{\pi}$ ne tend pas vers 0. L'estimateur est sans biais mais il n'est pas convergent.