

# INFÉRER DES ÉVÈNEMENTS DÉMOGRAPHIQUES DANS UNE POPULATION STRUCTURÉE

Stage de L3

---

Max Halford

28 août 2015

CMI SID

- **Lieu** - INSA et Institut Gulbenkian
- **Thème** - La génétique des populations
- **Notions** - modélisation mathématique, optimisation, génie logiciel
- **Superviseurs** - Olivier Mazet et Lounès Chikhi

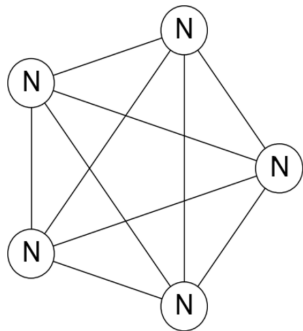
# GÉNÉTIQUE DES POPULATIONS

---



- On retrace la généalogie d'individus jusqu'à trouver l'ancêtre commun le plus récent
- Temps de coalescence pour deux individus =  $T_2$
- Le  $T_2$  est lié à la taille et à la structure d'une population

# POPULATIONS STRUCTURÉES



- La procréation se fait au sein de sous-populations appelées “îles”
- Les individus peuvent migrer d’île en île
- Le taux de migration peut changer au cours du temps

# MODÉLISATION MATHÉMATIQUE

---

- Trois états:
  - $s$ , les individus sont dans la même île
  - $d$ , les individus sont dans deux îles différentes
  - $c$ , Les individus ont coalescé
- Utilisation de la loi exponentielle car elle est continue et elle vérifie la propriété de non-vieillessement
- Les probabilités de passages d'un état à un autre dépendent de trois paramètres
  - $n$ , le nombre d'îles
  - $M$ , les taux de migrations
  - $T$ , les temps de changement du taux de migration

Outil stochastique permettant de calculer la probabilité de passer de l'état  $i$  à l'état  $j$  à l'instant  $t$ , qu'on note  $p_{ij}(t)$ . Ici,

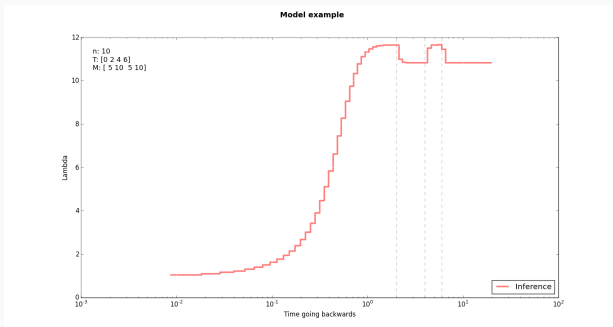
$$Q = \begin{pmatrix} -(M+1) & M & 1 \\ \frac{M}{n-1} & -\frac{M}{n-1} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$p_{ij}(t) = e^{tQ}, \forall t > 0.$$

$$\text{Taux de coalescence instantané (IICR)} = \frac{1-p_{13}(t)}{p_{11}(t)}$$



# PROBLÈME



- Pour une liste de temps et des paramètres on obtient une fonction en escalier.
- Comment faire l'inverse? Trouver les paramètres du modèle étant donnée une fonction.

# ALGORITHME GÉNÉTIQUE

---

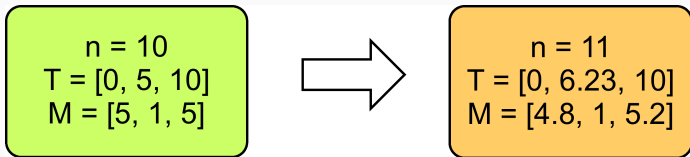
1. Produire un ensemble (population) de solutions aléatoires (individus)
2. Trier la population selon la performance (fitness) de chaque individu
3. Sélectionner un sous-ensemble de la population (parents)
4. Générer une nouvelle population par croisement ou par mutation des propriétés des individus (ADN)
5. Répéter l'étape 2 avec la nouvelle population

# MUTATION PAR LOI NORMALE

1. Choisir un sous-ensemble des paramètres
2. Pour chaque paramètre  $p$  du sous-ensemble
3. Considérer le nouveau paramètre comme une variable aléatoire qui suit une  $\mathcal{N}(p, \sigma^2)$

$\sigma^2$  détermine la taille de l'espace de recherche exploré.

Par exemple:



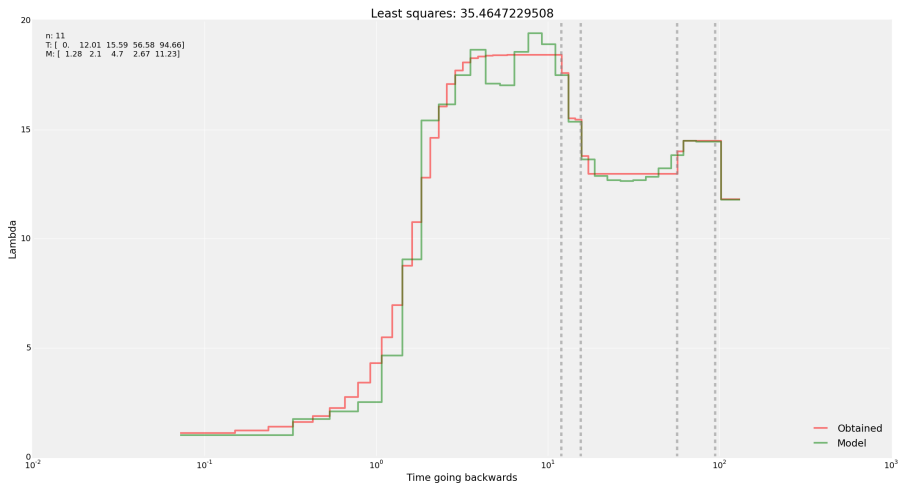
1. Choisir un échantillon de la population de taille  $k$
2. Garder le meilleur individu de l'échantillon
3. Répéter l'étape 1 avec ou sans remise

$k$  détermine la vitesse de convergence de l'algorithme.

- Nécessité de tester différentes typologies
- Génération de 240 histoires démographiques
- 24 heures de temps de calcul
- L'algorithme estime bien les scénarios basiques (8 dimensions ou moins)
- L'algorithme a du mal avec plus de 8 dimensions quand les changements de taux de migration sont rapprochés

# EXEMPLE 1 - 10 ÎLES ET 4 CHANGEMENTS

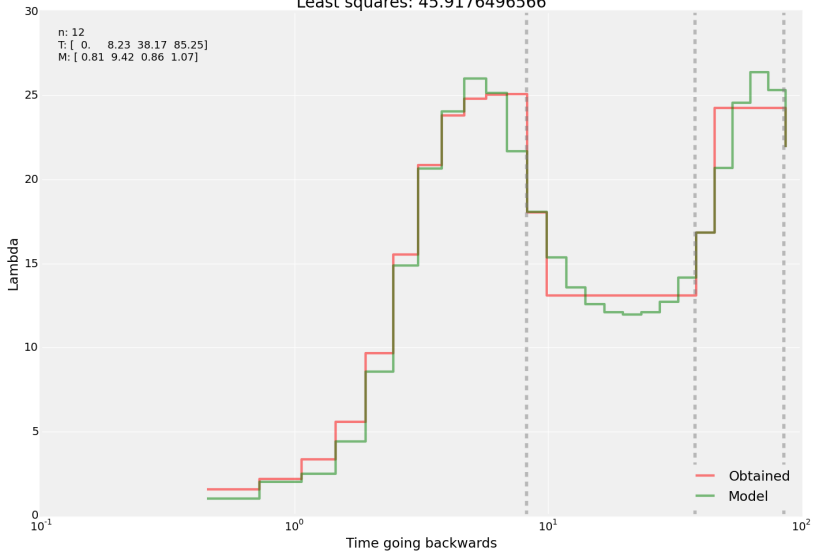
Structured model inference of lambda function



# EXAMPLE 2 - LI & DURBIN

## Structured model inference of lambda function

Least squares: 45.9176496566





# GÉNIE LOGICIEL

---

- 3 moyens d'installation
- Lecture de fichiers CSV et PSMC
- Utilisation à la ligne de commande
- Possibilité de relancer l'algorithme plusieurs fois
- Modification aisée du graphe de sortie
- Programmation objet = modulable
- Programmation fonctionnelle = lisible
- Documentation en ligne

- Vérification de l'installation
- Vérification des modules
  - Le modèle mathématique
  - Le tracage des graphes
  - L'algorithme génétique

<http://maxhalford.com/StSICMR-Inference/>

**MERCI POUR VOTRE ATTENTION**