Modèles stochastiques de populations structurées en génétique des populations neutre.

mohammed.sedki@u-psud.fr

Université Paris-Sud (Paris-Saclay), Inserm et Institut Pasteur

I. Plusieurs populations structurées

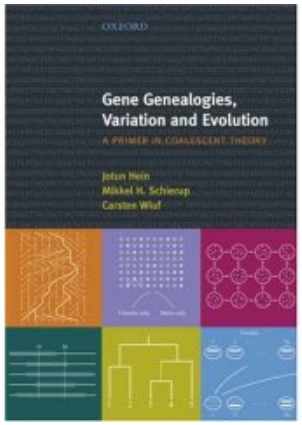
Motivation biologique

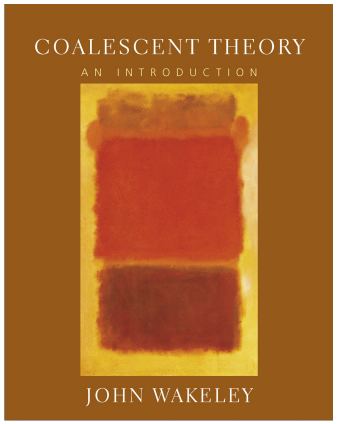
Pourquoi fait-on de la génétique des populations ?

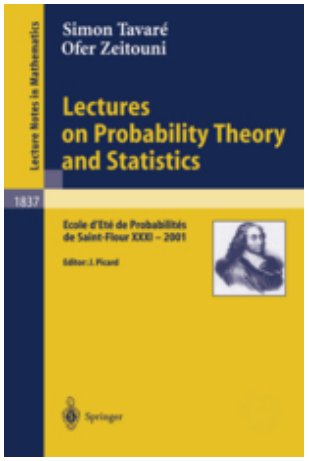
- Estimer les fréquences des allèles
- Déterminer la répartition des allèles dans les différentes populations
- Prédire et comprendre l'évolution des fréquences des allèles dans les populations sous l'effet de différents facteurs

☞Analyser l'effet des différentes forces évolutives (mutation, dérive, migration, sélection) sur l'évolution des fréquences des allèles dans le temps et l'espace.

La brique de base est simple à implémenter : le coalescent de Kingman







☞ Le but est de reconstruire des éléments de l'histoire des populations. Pour examiner la structure des données génétiques, ces méthodes utilisent l'arbre généalogique des gènes.

Un modèle est un scénario !

La formulation d'un modèle est contrainte par un scénario évolutif qui imite la réalité historique et démographique de l'espèce.

Un tel scénario résume l'histoire évolutive des populations par une suite d'événements démographiques depuis une population ancestrale.

Ces événements sont constitués de **divergences** avec ou sans **remises en contact**, des **migrations** et des **variations de tailles** entre les populations.

Le jeu de données : observation de l'espèce dans le présent!

Nos jeux de données sont constitués d'informations génétiques issues de plusieurs locus.

☞ Nous nous restreindrons à l'hypothèse d'indépendance. Cette hypothèse est justifiée dès que les locus sont suffisamment éloignés dans le génome.

Une classe de modèles

Ici, on s'intéresse à une classe de modèles probabilistes constitués d'événements **inter-populationnels** comme la **divergence**, l'**admixture** et la **migration**.

☞ Les modèles que nous étudions sont sous l'hypothèse de neutralité Kimura (1968, 1983). Cette hypothèse implique l'**absence d'effet de sélection**.

☞Le **polymorphisme** expliqué par les modèles évolutifs est le résultat de la **superposition des mutations** génétiques sur la généalogie des individus.

Cette classe de modèles est riche !!

L'inférence des paramètres du modèle permet de :

- Dater des divergences et des admixtures
- Quantifier des réductions ou des augmentations de tailles efficaces de populations.
- Inférer des taux de migration

Une procédure de choix de modèle (scénario démographique) permet:

- Déterminer de quelle source ancestrale provient une population récente.
- Décrire des voies d'invasion de populations.

L'exemple des orangs-outans (estimation)

- There are demographic evidences that orang-utan population sizes have collapsed

→ but what is the major cause of the decline, when did it start and how strong is it ?

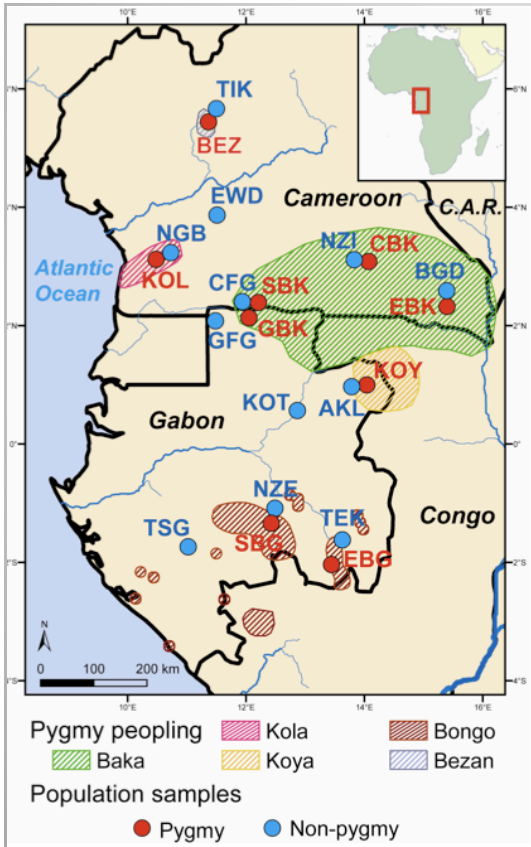
- Can population genetics help ?
  - Can we infer the time of the event ?
  - Can we infer the strength of the population size decrease ?

Deux conclusions

☞ La diversité génétique était plus faible chez les orangs-outans de Bornéo (Pongo pygmaeus) que chez ceux de Sumatra (Pongo abelii), bien que ceux de Bornéo soient six ou sept fois plus nombreux que ceux de Sumatra.

☞ On estime que ces deux espèces ont divergé il y a autour de 400,000 ans.

Un exemple d'étude sur les pygmées (choix de modèle)



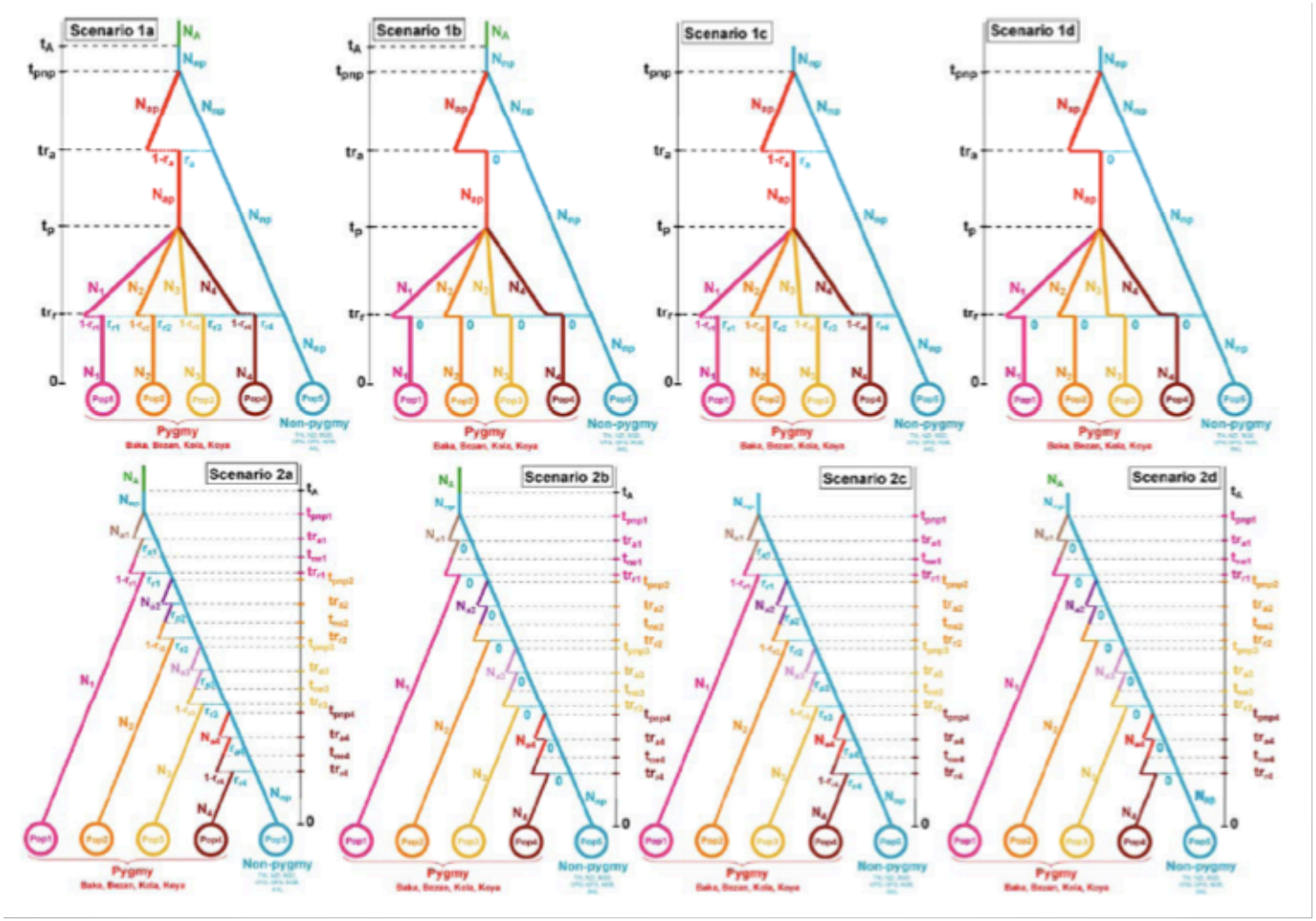
Crédit : Serge Bahuchet

604 individus, 12 populations non-pygmées, 9 populations pygmées, 28 marqueurs microsatellites

Verdu *et al.* (2009) *Current Biology* **19**: 312-318

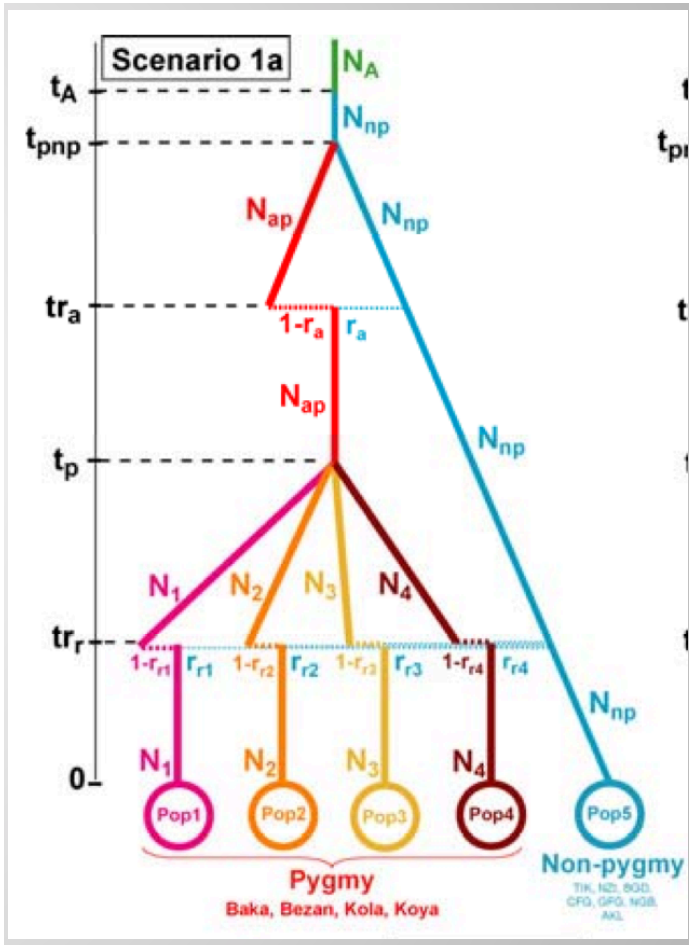
Les pygmées ont-ils une origine commune ? Y-a-t-il beaucoup d'échanges entre populations pygmées et non-pygmées ?

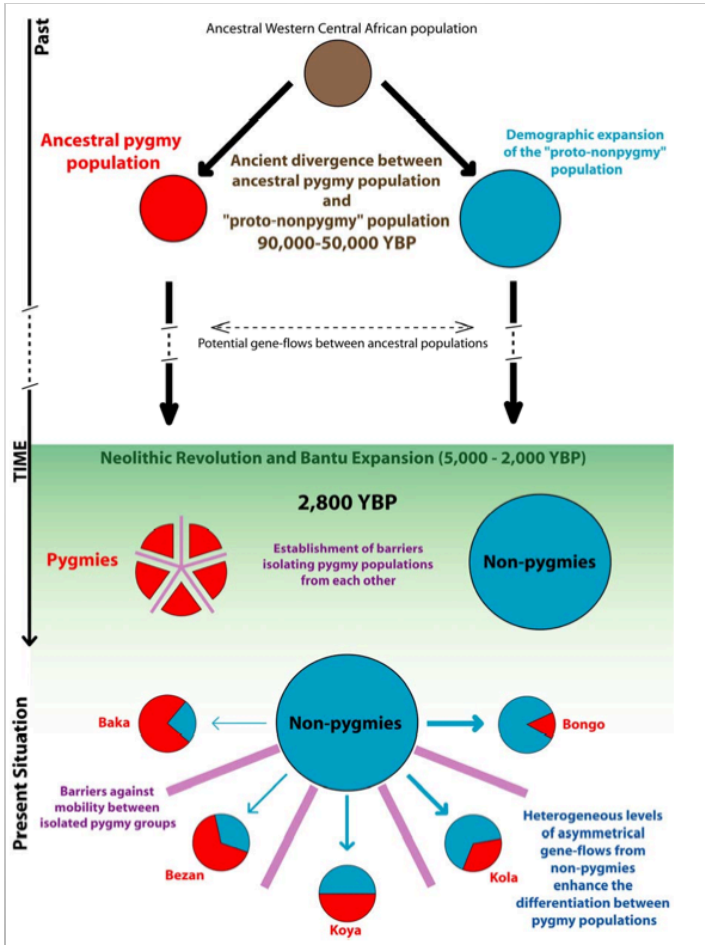
Les modèles en compétition



Conclusion de l'étude

☞ Le modèle ou le scénario retenu :





Les données

Un jeu de données

Le jeu de données est constitué de différents échantillons d'individus.

Chaque échantillon correspond à une **population** géographique.

Nous numérotons les populations de 1 à D et leur donnons les labels $Pop1$ à $PopD$.

Par population, la taille typique d'un échantillon varie entre une quarantaine et une centaine d'individus. La taille de l'échantillon issu de la population $Popi$ est notée n_i.

Un peu de vocabulaire

La plupart des espèces sont **diploïdes**.

Les individus portent l'information génétique nucléaire en double : une copie issue de la gamète maternelle, une copie issue de la gamète paternelle.

On peut donc assimiler un individu diploïde aux deux gamètes qui l'ont engendré, c'est-à-dire à deux individus haploïdes.

Nous considérerons donc que les individus sont **haploïdes**.

Un peu de vocabulaire (suite)

Pour chacun des individus, l'information génétique que l'on considère dans l'étude est limitée.

On ne s'intéresse qu'à quelques positions particulières du génome appelées **locus**.

À ces locus, la séquence d'ADN peut varier d'un individu à l'autre, à cause des **mutations** au cours de l'évolution de l'espèce.

On parle alors de **polymorphisme génétique**. Les différentes variantes s'appellent des **allèles** ou des états alléliques.

Un peu de vocabulaire (suite)

La constitution de notre jeu de données a nécessité de déterminer l'allèle que porte chacun des individus pour tous les locus de l'étude.

Il existe trois types de locus : **microsatellite**, **SNP** (Single Nucleotide Polymorphism) ou séquence.

☞ On sait faire quand il s'agît de microsatellite

Marqueurs microsatellites

☞ Une partie de l'ADN où un court motif (de 1 à 4 paires de base) est répété en de nombreux exemplaires.

☞ **Fort polymorphisme**, cette partie de l'ADN est **très utilisée** en génétique des populations.

Justifions un peu le choix du coalescent

Certains modèles (Wright-Fisher et Moran) proposent de simuler l'évolution de la population entière, du passé au présent, puis d'échantillonner la dernière génération.

Trop lent pour une population de grande taille, la simulation suivant ce type de modèles est très lente.

☞ **On s'intéresse seulement à l'évolution des ascendants des individus de notre échantillon en remontant le temps.**

Le coalescent de Kingman (Kingman (1982), Tajima, Tavaré...)

☞ La généalogie d'un échantillon d'individus est représenté par un dendrogramme.

☞ On génère des lignées ancestrales jusqu'à l'ancêtre commun le plus récent (MRCA en anglais).

☞ Un évènement de coalescence se produit lorsque les lignées de deux individus se rejoignent en un noeud du dendrogramme.

☞ La généalogie d'un échantillon de k individus est donc composée de $k-1$ évènements de coalescence.

Petit rappel

Soient $T_k, \ldots, T_2$ les durées entre les événements de coalescences successifs.

La loi de la généalogie de k individus est entièrement caractérisée par la loi du choix des lignées à chaque évènement de coalescence et la loi des durées entre évènements $T_k, \ldots, T_2$.

☞ Pour le coalescent de Kingman, les **durées entres évènements** de coalescences $T_k, \ldots, T_2$ sont **indépendantes** et T_k suit la loi **exponentielle** de paramètre $k\big(k-1\big)/2$.

Simuler un coalescent

Une unité de temps coalescent s'interprète comme Ne générations, taille efficace de la population. Lorsque le temps est à l'échelle naturelle, le taux de coalescence dans la généalogie est linéaire en Ne.

Tantque $k \geq 2$ **faire**

1) Simuler le temps inter-coalescent T_k suivant une loi exponentielle de paramètre $\frac{k\left(k-1\right)}{2Ne}$.

2) Augmenter les longueurs des k lignées de T_k.

3) Parmi les k lignées, choisir aléatoirement deux lignées à regrouper pour former un noeud du dendrogramme.

4) $k \leftarrow k-1$.

Fin tantque

Simuler la généalogie de plusieurs populations structurées

Plusieurs populations structurées

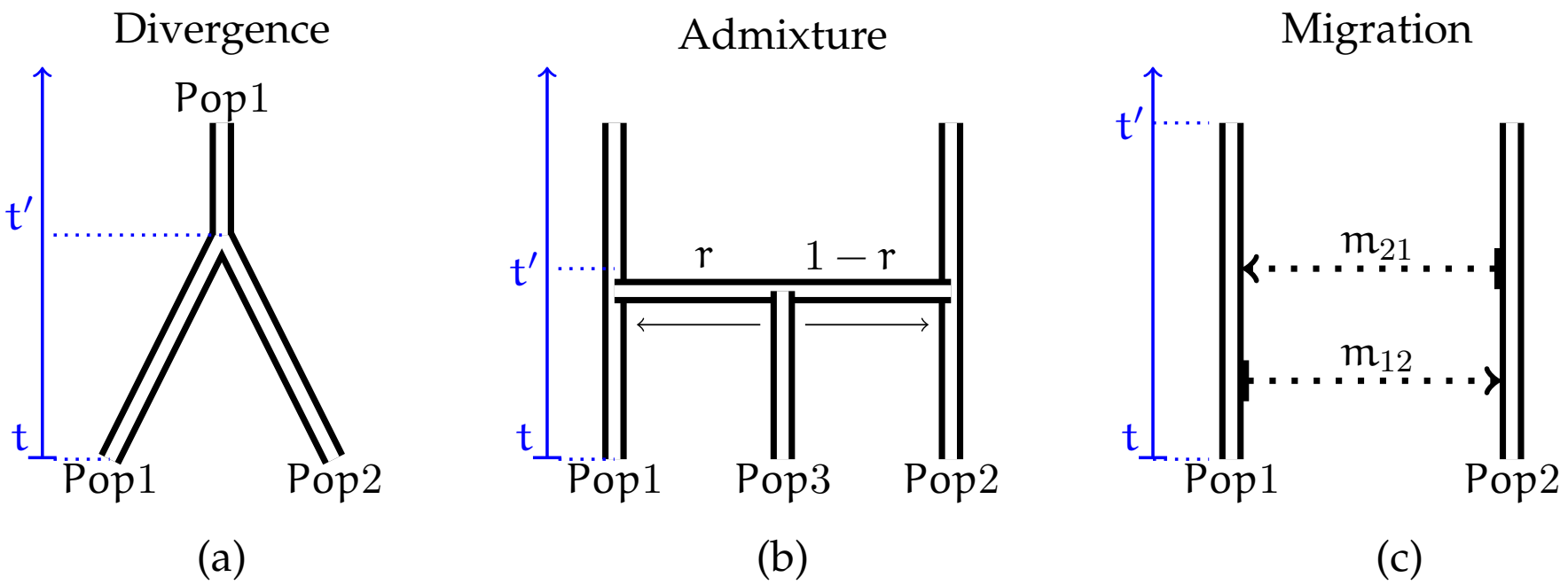
On s'intéresse à la loi d'une généalogie pour un scénario évolutif dont la structure géographique est gouvernée par des événements inter-populationnels.

On combine ces événements avec le coalescent de Kingman qui décrit la généalogie intra-populationel.

Trois types d'événements inter-populationnels :

- **La divergence** est la fusion de deux populations.
- **L'admixture** est le partage d'une population en deux partie à l'instant de l'événement.
- **La migration** autorise le déplacement des lignées d'une population à l'autre sur une période donnée.

Les trois types d'événements



Un algorithme générique

Trier les événements inter-populationnels du plus récent au plus ancien.

Pour t allant de l'événement le plus récent au plus ancien **faire**

1) Simuler les généalogie intra-populationnel: un coalescent de Kingman indépendant par population jusqu'à t ou combiner le coalescent de Kingman avec migration dans le cas d'une migration.

2) Appliquer l'événement inter-populationnels instantané à la date t.

Fin pour

Simuler un (ou des) coalescent(s) de Kingman (avec migrations) sur la (les) dernière(s) population(s) jusqu'au MRCA.

L'évolution de la généalogie intra-populations entre deux dates (notées t et t' où $t' > t$.

Simuler T_k suivant une loi exponentielle de paramètre $k\big(k-1\big)/2Ne$.
Tantque $\big(t + T_k\big) \leq t'$ **faire**

1) Augmenter les longueurs des k lignées d'une longueur T_k.

2) Choisir aléatoirement parmi les k lignées, deux lignées à regrouper pour former un noeud du dendrogramme.

3) $k \leftarrow k - 1$.

4) Simuler la durée inter-coalescence T_k suivant une loi exponentielle de paramètre $k\big(k-1\big)/2Ne$.

Fin tantque

Si $\big(t + T_k\big) > t'$ **alors**
Augmenter les lignées restantes jusqu'à la hauteur t'.
Fin si

À l'instant de la divergence, les lignées présentes dans les deux populations sont regroupées pour former une seule population.

À l'instant de l'admixture, l'échantillon ancestral de $Pop3$ est partagé sur les deux autres populations ainsi: une lignée de la population $Pop3$ est envoyée dans $Pop1$ avec probabilité r et dans $Pop2$ avec probabilité $1-r$, où r est un paramètre du modèle appelé taux d'admixture.

En cas de présence d'un changement de taille efficace dans la population à une date : changer l'échelle de temps après cette date (remplacer Ne par Ne').

La migration est paramétrée par les taux de migration de populations i vers j : m_{ij}.

Pour i $=1 \to D$ **faire**

1) Associer une horloge exponentielle de paramètre $1/Ne_i$ pour chaque couple d'individus de la population i qui correspond à une coalescence potentielle.

2) Associer $D-1$ horloges exponentielles de paramètres $m_{ij}, 1 \leq j \neq i \leq D$ pour chaque individu de la population i qui correspondent à des migrations potentielles.

Fin pour

Parmi toutes les horloges en compétition, celle qui sonne en premier gagne. Si cette horloge correpond à un couple d'individus, on fait coalescer ces deux individus. Si c'est l'horloge d'un seul individu, et celle-ci est de paramètre m_{ij}, on déplace la lignée de cet individu des populations i vers j.

Générer les données conditionnellement à la généalogie
Processus mutationnels

1. Jeter les mutations sur la généalogie

Le taux de mutation par unité de temps naturel et par individu diploïde est le paramètre μ.

Conditionnellement à une généalogie, les positions des mutations sont données par un processus ponctuel de Poisson d'intensité $\mu/2$ sur le dendrogramme.

Sur une branche de longueur t, le nombre N de mutations suit une loi de Poisson de paramètre $\mu t/2$, et les N mutations sont uniformément réparties sur cette branche.

2. Générer les données suivant un modèle mutationnel

Deux modèles mutationnels : SMM (Stepwise Mutation Model) et GSM (Generalized Stepwize Mutation Model)

Les chaînes de Markov associées aux modèles SMM et GSM sont des marches aléatoires symétriques sur un intervalle de nombres entiers $[\![a; b]\!]$ de $\mathbb{N}$.

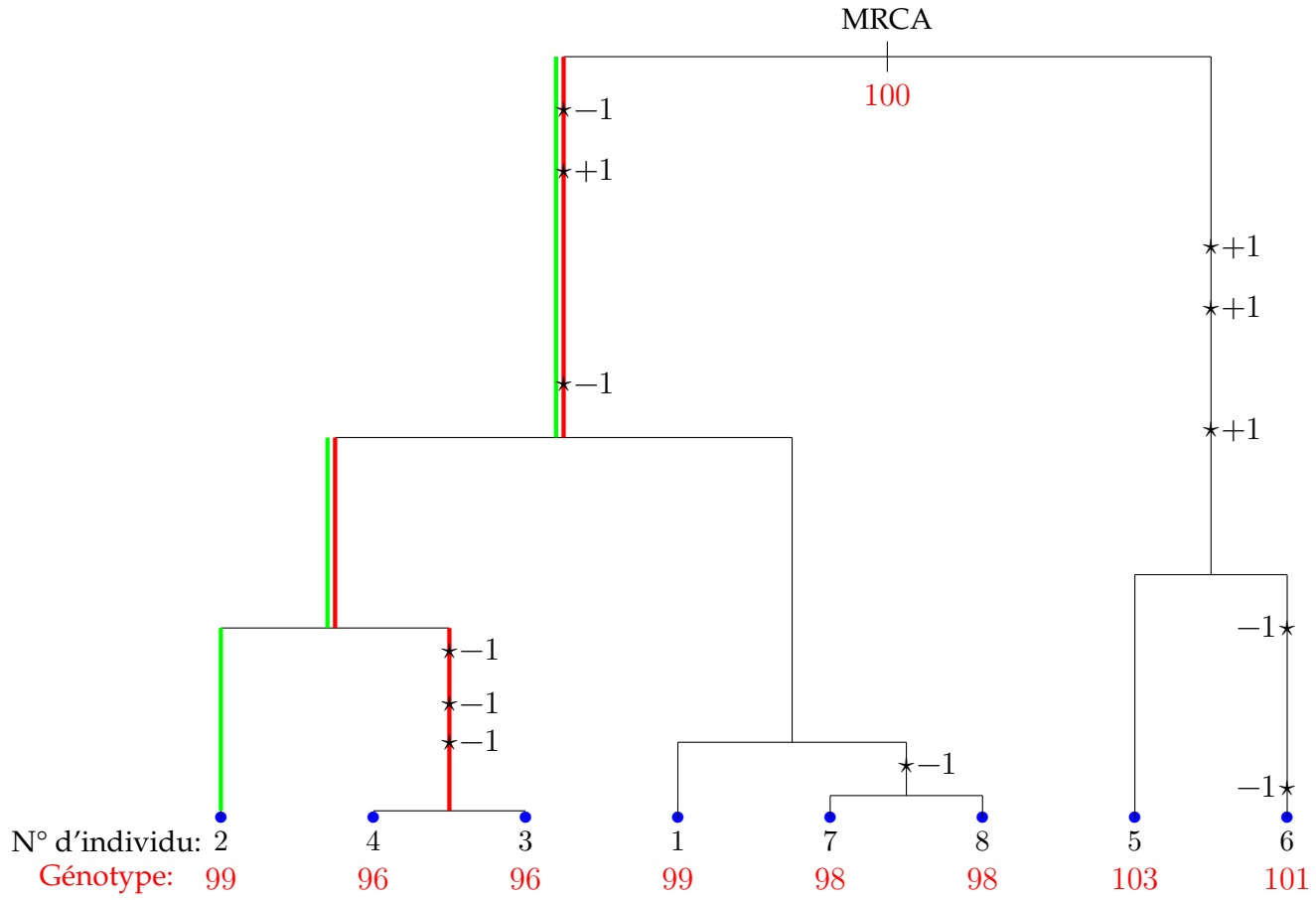
Modèle GSM : modifier le locus d'une longueur $\pm mG$, où m est la longueur connue du motif répété, G est une variable aléatoire de loi géométrique de paramètre p et un signe aléatoire (avec probabilités 1/2 et 1/2 respectivement).

En pratique, le paramètre p est de l'ordre de 0.2.

Modèle SMM : une mutation revient à diminuer ou augmenter (avec probabilités 1/2 et 1/2 respectivement) le locus d'une longueur de m paires de base où m est la longueur du motif répété.

Il arrive qu'en appliquant les mutations, le génotype dépasse les bornes a et b de l'ensemble des états alléliques : troncature aux bornes.

Pour simuler les génotypes de l'échantillon en un locus donné, il suffit de faire évoluer le génotype du MRCA le long de la généalogie jusqu'au présent en appliquant les mutations.


MRCA
100
⋆−1
⋆+1
⋆−1
⋆+1
⋆+1
⋆+1
⋆−1
⋆−1
⋆−1
⋆−1
−1⋆
−1⋆
N° d'individu: 2 4 3 1 7 8 5 6
Génotype: 99 96 96 99 98 98 103 101

La vraisemblance

Bilan

Chaque modèle est caractérisé par un ensemble de paramètres $\boldsymbol{\theta}$ **historiques** (temps de divergence, temps d'admixture, ...), **démographiques** (tailles efficaces, taux d'admixture, taux de migrations, ...) et **génétiques** (taux de mutation, ...).

Le but est d'estimer ces paramètres à partir d'un jeu de données de polymorphisme (échantillon génétique) $\mathbf{x}$ observé au temps présent.

Problème : la plupart du temps, on ne sait pas calculer la vraisemblance de données de polymorphisme $f(\mathbf{x}|\boldsymbol{\theta})$.

Notons $f_{\boldsymbol{\theta}}(\mathcal{G})$ la densité de la loi de la généalogie de gènes par rapport à une mesure de référence $\mathrm{d}\mathcal{G}$.

Notons $f_{\boldsymbol{\theta}}(\mathcal{M}|\mathcal{G})$ la densité du processus mutationnel $\mathcal{M}$ sachant la généalogie $\mathcal{G}$.

Vraisemblance :

$$\ell(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i\in\{locus\}} \int_{\mathcal{M}_i\to\mathbf{x}_i} f_{\boldsymbol{\theta}}\big(\mathcal{M}_i\big|\mathcal{G}_i\big) f_{\boldsymbol{\theta}}\big(\mathcal{G}_i\big)\,\mathrm{d}\mathcal{G}_i\,\mathrm{d}\mathcal{M}_i, \qquad (1)$$

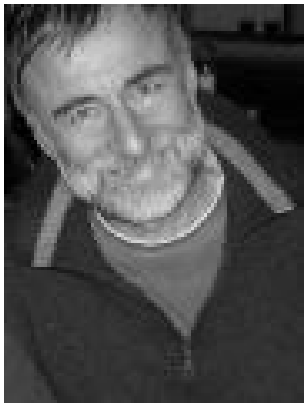
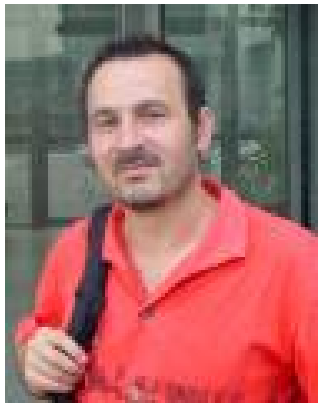
où $\mathbf{x}_i$ est l'ensemble des données au locus i et $\mathcal{M}_i \to \mathbf{x}_i$ désigne l'ensemble des génotypes sur le dendrogramme dont les feuilles correspondent à l'échantillon observé.

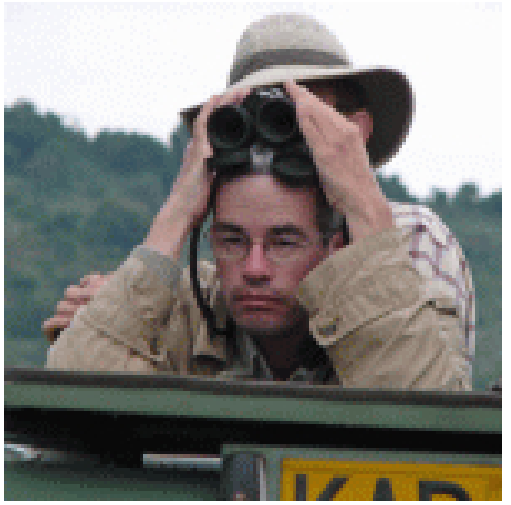
Cette vraisemblance ne se calcule pas facilement. L'intégrale précédente est sur l'espace des couples $(\mathcal{G}_i, \mathcal{M}_i)$ compatibles avec l'échantillon $\mathbf{x}_i$.

Cet espace est de très grande dimension et comporte des directions discrètes comme les génotypes des ancêtres et des parties continues comme les différentes hauteurs dans la généalogie.

En dépit de la simplicité du coalescent de Kingman et du processus mutationnel, on ne peut espérer aucune simplification formelle dans cette intégrale.

Merci à



Jean-Marie Cornuet, Arnaud Estoup, Raphaël Leblois et François Rousset

II. Inférence

Méthodes bayésiennes approchées

Approximate Bayesian Computation

Pré-requis de statistique bayésienne

On se place dans un contexte paramétrique : le vecteur des observations $\mathbf{x} \sim f(\mathbf{x}|\boldsymbol{\theta})$ où $\boldsymbol{\theta} \in \Theta$ **est un espace de dimension finie.**

L'information fournie par l'observation $\mathbf{x}$ sur $\boldsymbol{\theta}$ est contenue dans la densité $f(\mathbf{x}|\boldsymbol{\theta})$, que l'on représente classiquement sous la forme inversée de *vraisemblance*,

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}), \tag{2}$$

pour traduire qu'il s'agit d'une fonction de $\boldsymbol{\theta}$, qui est *inconnu*, dépendant de la valeur observée $\mathbf{x}$.

Loi *a priori*

Le paramètre inconnu $\boldsymbol{\theta}$ n'est pas considéré comme inconnu et déterministe, mais comme une **variable aléatoire**.

On considère que l'*incertitude* sur le paramètre $\boldsymbol{\theta}$ d'un modèle peut être décrite par une distribution de *probabilité* π sur Θ, appelée *distribution a priori*.

Ce qui revient à supposer que $\boldsymbol{\theta}$ est distribué suivant $\pi(\boldsymbol{\theta})$, $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$, "avant" que $\mathbf{x}$ ne soit généré suivant $f(\mathbf{x}|\boldsymbol{\theta})$, le conditionnement implicite dans cette notation prenant alors tout son sens.

Loi *a posteriori*

Par application directe du théorème de Bayes, la loi de $\boldsymbol{\theta}$ conditionnelle à $\mathbf{x}$, $\pi(\boldsymbol{\theta}|\mathbf{x})$, appelée *distribution a posteriori*, est définie par

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{\ell(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\int_\Theta \ell(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \ .$$

Cette densité est centrale pour l'inférence bayésienne en ce qu'elle suffit à déterminer les procédures de décision et, par extension, à conduire toute inférence liée à $\boldsymbol{\theta}$.

Estimation ponctuelle

On peut comparer les approximations d de $\boldsymbol{\theta}$ au moyen d'une fonction de coût, $L(d, \boldsymbol{\theta})$.

Une fois construite la loi a posteriori $\pi(\boldsymbol{\theta}|\mathbf{x})$, les approximations d ont un coût moyen égal à $\mathbb{E}^{\pi}(L(d, \boldsymbol{\theta})|\mathbf{x})$

L'approximation ou estimation optimale est celle qui minimise cette erreur. Un estimateur bayésien est

$$\delta(\mathbf{x}) = \arg\min_{d \in \Theta} \mathbb{E}^{\pi}[L(d, \boldsymbol{\theta})|\mathbf{x}] \,.$$

Un exemple jouet

Une fonction de perte par défaut est la fonction de perte quadratique :

$$L(d, \boldsymbol{\theta}) = (d - \boldsymbol{\theta})^2 \, .$$

Dans ce cas, l'estimateur bayésien est, si elle existe, l'espérance de la loi a posteriori $\delta(\mathbf{x}) = \mathbb{E}^{\pi}[\boldsymbol{\theta}|\mathbf{x}]$.

Exemple : Si $x \sim \mathcal{N}_1(\theta, 1)$ *et* $\theta \sim \mathcal{N}_1(0, 10)$,

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta) \propto \exp\left(-0.5\left\{.1\theta^2 + (\theta - x)^2\right\}\right) \, ,$$

ce qui équivaut à la loi $\theta|x \sim \mathcal{N}(10x/11, 10/11)$.
L'espérance a posteriori de θ *est donc* $10x/11$.

Régions de crédibilité (analogue de l'IC)

La connaissance de la distribution a posteriori permet la détermination des *régions de confiance* sous la forme de régions de plus forte densité a posteriori (*Highest Posterior Density*, HPD), c'est-à-dire des régions de la forme

$$\{\boldsymbol{\theta}; \pi(\boldsymbol{\theta}|\mathbf{x}) \geq k\},$$

dans le cas multidimensionnel comme dans le cas unidimensionnel.

La motivation conduisant à cette forme **de région de crédibilité** est que ces régions sont de volume minimal à un niveau nominal donné.

Un exemple jouet (suite)

Pour

$$\theta|x \sim \mathcal{N}_1(10x/11, 10/11)\,,$$

la région de crédibilité au α *est de la forme*

$$C_\alpha = \{\theta\,;\, \pi(\theta|x) \geq k\} = \{\theta\,;\, |\theta - 10x/11| \leq k'\}$$

avec k *et* k' *choisis de manière à ce que* $\pi(C_\alpha|\mathbf{x}) = \alpha$.

On obtient

$$(10x/11 - q_{1-\alpha/2}\sqrt{10/11}, 10x/11 + q_{1-\alpha/2}\sqrt{10/11})$$

où $q_{1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de la normale centrée réduite.*

Choix bayésien de modèles

Considérons deux modèles dénotés $\mathfrak{M}_1$ et $\mathfrak{M}_2$, où $(i = 1, 2)$

$$\mathfrak{M}_i : \mathbf{x} \sim f_i(\cdot|\boldsymbol{\theta}_i), \boldsymbol{\theta}_i \in \Theta_i, \boldsymbol{\theta}_i \sim \pi_i(\boldsymbol{\theta}_i)\,.$$

Les tests d'hypothèses correspondent aussi à des lois a priori dégénérées sur certaines composantes de $\boldsymbol{\theta}$.

L'ensemble des modèles d'une loi de probabilité a priori : $\mathbb{P}(\mathfrak{M}_1)$ et $\mathbb{P}(\mathfrak{M}_2)$.

Le choix de modèle bayésien est alors basé sur la loi a posteriori des différents modèles,

$$\mathbb{P}(\mathfrak{M}_i|\mathbf{x}) \propto \mathbb{P}(\mathfrak{M}_i) \int_{\Theta_i} f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_i \,.$$

Notons que cette distribution a posteriori est sensible au choix des lois a priori des paramètres des modèles.

Cette représentation impose l'utilisation de véritables lois de probabilités $\pi_i(\boldsymbol{\theta}_i)$, excluant l'emploi de lois impropres.

Difficultés

Le calcul explicite de la constante de normalisation de $\pi(\boldsymbol{\theta}|\mathbf{x})$ n'est pas possible.

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{\ell(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})}{\int_\Theta \ell(\boldsymbol{\theta}|\mathbf{x})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \ .$$

Idem pour la vraisemblance intégrée élément central de la loi a posteriori l'espace des modèles.

$$\mathbb{P}(\mathfrak{M}_i|\mathbf{x}) \propto \mathbb{P}(\mathfrak{M}_i) \int_{\Theta_i} f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_i \ .$$

On a recours à des méthodes de simulation.

Une situation typique :

Soit $(\mathbf{X}, \mathcal{B}(\mathbf{X}), \Pi)$ un espace de probabilité.

(A1) $\Pi \ll \mu$ et $\Pi(dx) = \pi(x)\mu(dx)$ (une densité par rapport μ).

(A2) On connait π à une constante (de normalisation) près :

- $\pi(x) = \dfrac{\tilde{\pi}(x)}{\displaystyle\int \tilde{\pi}(x)\mu(dx)}$;
- $\tilde{\pi}$ est accessible;
- Le calcul $\displaystyle\int \tilde{\pi}(x)\mu(dx) < \infty$ est impossible.

Le problème typique de la statistique bayésienne :

Problème: Soit une fonction test h Π-mesurable, approcher

$$\Pi(h) = \int h(x)\pi(x)\mu(dx) = \frac{\int h(x)\tilde{\pi}(x)\mu(dx)}{\int \tilde{\pi}(x)\mu(dx)}$$

(A3) Le calcul de $\int h(x)\tilde{\pi}(x)\mu(dx)$ est impossible.

La densité *a posteriori*: $\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Méthodes Monte-Carlo (principe)

☞ Générer $x_1, \ldots, x_N$ iid à partir de Π et estimer $\Pi(h)$ par

$$\widehat{\Pi}_N^{\mathrm{mc}}(h) = N^{-1} \sum_{i=1}^{N} h(x_i).$$

$$\widehat{\Pi}_N^{\mathrm{mc}}(h) \longrightarrow_{as} \Pi(h)$$

If $\Pi(h^2) = \int h^2(x)\pi(x)\mu(dx) < \infty$,

$$\sqrt{N}(\widehat{\Pi}_N^{\mathrm{mc}}(h) - \Pi(h)) \longrightarrow_{\mathcal{L}} \mathcal{N}(0, \Pi((h - \Pi(h))^2)).$$

☞ Souvent, on ne sait pas simuler suivant Π directement !

Monte-Carlo par chaînes de Markov (MCMC)

☞ Générer une trajectoire $x^{(1)}, \ldots, x^{(T)}$ d'une chaîne de Markov $(x_t)_{t\in\mathbb{N}}$ avec Π et estimer $\Pi(h)$ par

$$\widehat{\Pi}_N^{\mathrm{mcmc}}(h) = N^{-1} \sum_{i=T-N+1}^{T} h\left(x^{(i)}\right).$$

☞ La convergence vers la loi stationnaire peut être lente !

Un "Metropolis-Hastings" générique

On veut simuler suivant une loi *cible* de densité π :

Un état initial : Choisir un $x^{(0)}$ arbitraire

Itération t:

1. Conditionnellement à $x^{(t-1)}$, générer $\tilde{x} \sim q(x^{(t-1)}, x)$
2. Calculer

$$\rho(x^{(t-1)}, \tilde{x}) = \min\left(\frac{\pi(\tilde{x})/q(x^{(t-1)}, \tilde{x})}{\pi(x^{(t-1)})/q(\tilde{x}, x^{(t-1)})}, 1\right)$$

3. Avec probabilité $\min(\rho(x^{(t-1)}, \tilde{x}), 1)$, on accepte $\tilde{x}$ et on a $x^{(t)} = \tilde{x}$; sinon, rejet de $\tilde{x}$, et dans ce cas $x^{(t)} = x^{(t-1)}$.

☞ **Cas où q est associé à une marche aléatoire**

Illustration (si le temps le permet)

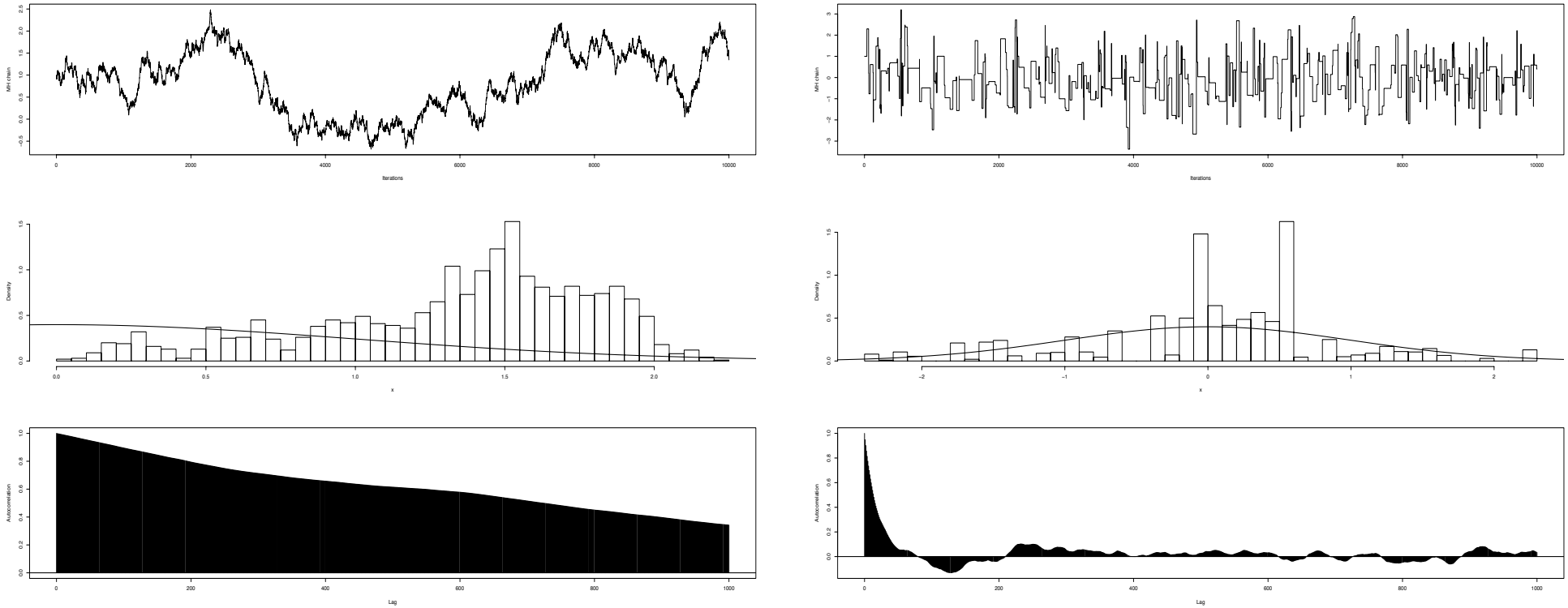
On *vise* la de densité d'une v.a de loi $\mathcal{N}(0,1)$.

On fait appel à un Metropolis-Hastings avec marche aléatoire gaussienne, *i.e.*

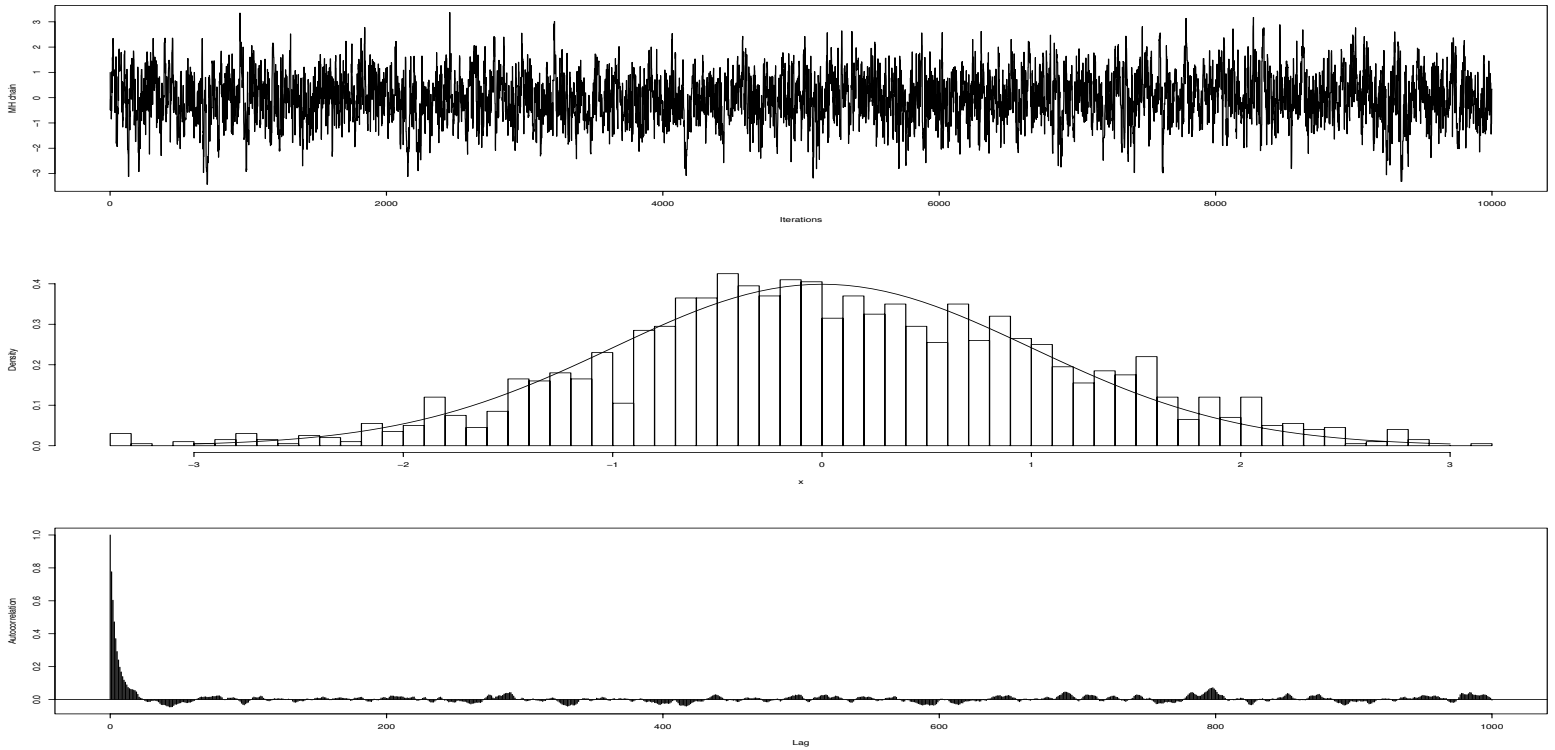
$$\tilde{x}|x^{(t-1)} \sim \mathcal{N}\left(x^{(t-1)}, \sigma^2\right) ,$$

$$q_{RW}(\tilde{x} - x^{(t-1)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}(\tilde{x} - x^{(t-1)}) ,$$

☞ Les performances dépendent du choix de σ^2 !!



(À gauche) $\sigma^2 = 10^{-4}$, *(à droite)* $\sigma^2 = 10^3$. *Haut*: une suite de $10,000$ itérations échantillonnée à chaque 10 itérations. *Au milieu*: histogramme des $2,000$ dernières itérations comparé à la vraie densité. *En bas*: auto-corrélations empiriques.


MH chain
Iterations
Density
x
Autocorrelation
Lag

Importance sampling (si le temps le permet)

Soit Q une mesure de probabilité sur $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$. On suppose que $\Pi \ll Q$, $Q \ll \mu$ et $Q(dx) = q(x)\mu(dx)$:

$$\Pi(h) = \int h(x)\{\pi/q\}(x)q(x)\mu(dx).$$

☞ Générer $x_1, \ldots, x_N$ iid suivant Q, appelée, *"proposal"* et estimer $\Pi(h)$ par

$$\widehat{\Pi}^{\text{is}}_{Q,N}(h) = N^{-1} \sum_{i=1}^{N} h(x_i)\{\pi/q\}(x_i).$$

$$\widehat{\Pi}^{\text{is}}_{Q,N}(h) \longrightarrow_{as} \Pi(h).$$

Si $Q\Big(\big(h\pi/q\big)^2\Big) < \infty$,

$$\sqrt{N}\Big(\widehat{\Pi}^{\text{is}}_{Q,N}(h) - \Pi(h)\Big) \longrightarrow_{\mathcal{L}} \mathcal{N}\left(0, Q\Big(\big(h\pi/q - \Pi(h)\big)^2\Big)\right).$$

Ça ne marche pas quand la constante de normalisation de π est inconnue !!

On fait appel à l'estimateur IS auto-normalisé,

$$\widehat{\Pi}^{\text{snis}}_{Q,N}(h) = \left(\sum_{i=1}^{N}\{\pi/q\}(x_i)\right)^{-1}\sum_{i=1}^{N}h(x_i)\{\pi/q\}(x_i).$$

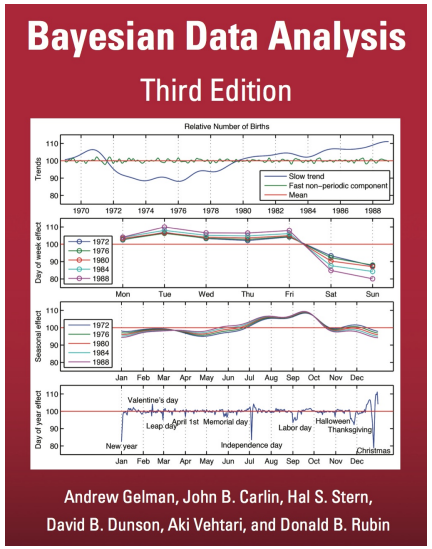
Importance sampling auto-normalisé

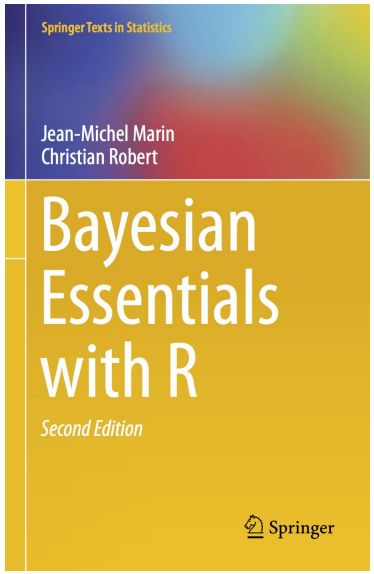
$$\widehat{\Pi}^{\text{snis}}_{Q,N}(h) \longrightarrow_{as} \Pi(h).$$

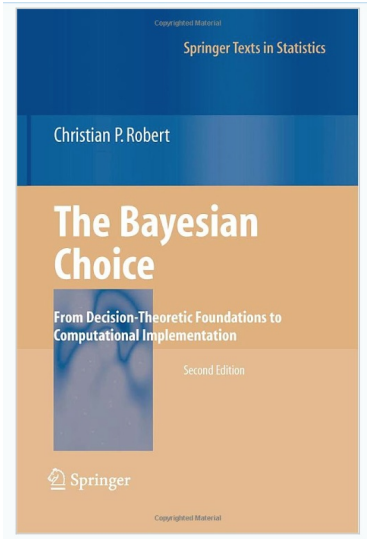
If $\Pi\Big((1+h^2)\big(\pi/q\big)^2\Big) < \infty$,

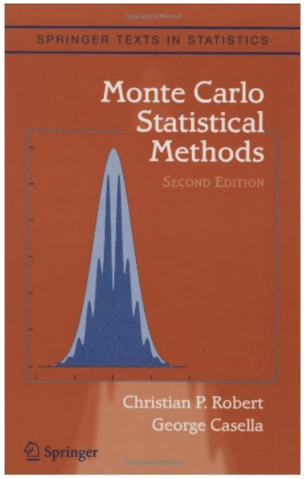
$$\sqrt{N}\Big(\widehat{\Pi}^{\text{snis}}_{Q,N}(h) - \Pi(h)\Big) \longrightarrow_{\mathcal{L}} \mathcal{N}\left(0, Q\Big(\big(\pi/q\big)^2\big(h-\Pi(h)\big)^2\Big)\right).$$

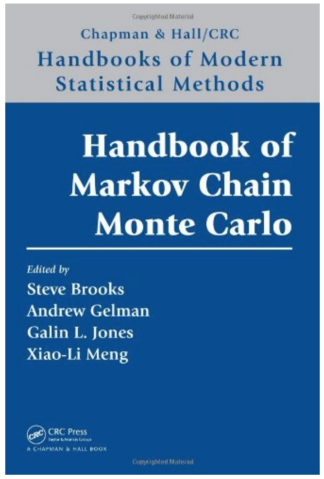
☞ **Le choix de Q est important pour les deux versions**


Bayesian Data Analysis
Third Edition
Relative Number of Births
Trends
Slow trend
Fast non−periodic component
Mean
Day of week effect
Seasonal effect
Day of year effect
Valentine's day
April 1st
Memorial day
Leap day
Labor day
Halloween
Thanksgiving
New year
Independence day
Christmas
Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin


Springer Texts in Statistics
Jean-Michel Marin
Christian Robert
Bayesian Essentials with R
Second Edition
Springer


Springer Texts in Statistics
Christian P. Robert
The Bayesian Choice
From Decision-Theoretic Foundations to Computational Implementation
Second Edition
Springer


SPRINGER TEXTS IN STATISTICS
Monte Carlo Statistical Methods
Second Edition
Christian P. Robert
George Casella
Springer


Chapman & Hall/CRC
Handbooks of Modern Statistical Methods
Handbook of Markov Chain Monte Carlo
Edited by
Steve Brooks
Andrew Gelman
Galin L. Jones
Xiao-Li Meng
CRC Press

Approximate Bayesian Computation (début)

La vraisemblance: $f(\mathbf{y}|\boldsymbol{\theta})$
Loi *a priori* sur $\boldsymbol{\theta}$: $\pi(\boldsymbol{\theta})$

Supposons que $\mathbf{y}$ est à valeurs dans un ensemble dénombrable noté $\mathcal{D}$.

Likelihood free rejection sampling 1 (Tavaré et al. (1997) Genetics)

1) Initialisation $i = 1$,

2) Générer $\boldsymbol{\theta}'$ suivant la loi à *a priori* $\pi(\cdot)$,

3) Generate $\mathbf{z}$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$,

4) If $\mathbf{z} = \mathbf{y}$, set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$ and $i = i + 1$,

5) If $i \leq N$, return to **2)**.

☞ On obtient un échantillon $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N)$ iid suivant la loi *a posteriori*.

La preuve (triviale):

$$f(\boldsymbol{\theta}_i) \propto \sum_{\mathbf{z}\in\mathcal{D}} \pi(\boldsymbol{\theta}_i) f(\mathbf{z}|\boldsymbol{\theta}_i) \mathbb{I}_{\mathbf{y}}(\mathbf{z}) \,,$$

$$f(\boldsymbol{\theta}_i) \propto \pi(\boldsymbol{\theta}_i) f(\mathbf{y}|\boldsymbol{\theta}_i) \,,$$

$$f(\boldsymbol{\theta}_i) = \pi(\boldsymbol{\theta}_i|\mathbf{y}) \,.$$

ABC pour l'estimation des paramètres (principe)

Likelihood free rejection sampling 2

(Pritchard et al. (1999) Mol. Biol. Evol.)

1) Set $i = 1$,

2) Generate $\boldsymbol{\theta}'$ from the prior distribution $\pi(\cdot)$,

3) Generate $\mathbf{z}$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$,

4) If $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \epsilon$, set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$ and $i = i + 1$,

5) If $i \leq N$, return to **2)**.

Dans un premier temps, on considère $\eta(\mathbf{z}) = \mathbf{z}$. ρ est la distance euclidienne.

ABC pour l'estimation des paramètres (idée)

L'algorithme précédent génère suivant la marginale en $\mathbf{z}$ de :

$$\pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z})}{\int_{A_{\epsilon,\mathbf{y}} \times \Theta} \pi(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) \mathrm{d}\mathbf{z} \mathrm{d}\boldsymbol{\theta}} \,,$$

- $\epsilon > 0$ un seuil,
- $\mathbb{I}_B(\cdot)$ la fonction indicatrice de l'ensemble B,
- $A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \epsilon\}$.

Pour ϵ suffisamment petit et un bon choix de η, nous avons l'approximation

$$\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}) = \int \pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) \mathrm{d}\mathbf{z} \approx \pi(\boldsymbol{\theta}|\mathbf{y}) \,.$$

Une approximation par convolution

$$\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}) = \frac{\int \pi(\boldsymbol{\theta}|\mathbf{z}) K_\epsilon(\mathbf{y},\mathbf{z}) \mathrm{d}\mathbf{z}}{\int \pi(\boldsymbol{\theta}|\mathbf{z}) K_\epsilon(\mathbf{y},\mathbf{z}) \mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta}} \,.$$

Une version MCMC de ABC

Likelihood free MCMC sampler (Majoram et al. (2003) PNAS)

1) Use the likelihood free rejection sampling to get a realization $(\boldsymbol{\theta}^{(0)}, \mathbf{z}^{(0)})$ from the ABC target distribution $\pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$,

2) Set $t = 1$,

3) Generate $\boldsymbol{\theta}'$ from the Markov kernel $q\left(\cdot|\boldsymbol{\theta}^{(t-1)}\right)$,

4) Generate $\mathbf{z}'$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$,

5) Generate u from $\mathcal{U}_{[0,1]}$,

6) If $u \leq \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t-1)}q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})}\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}')$,
set $(\boldsymbol{\theta}^{(t)}, \mathbf{z}^{(t)}) = (\boldsymbol{\theta}', \mathbf{z}')$ else $(\boldsymbol{\theta}^{(t)}, \mathbf{z}^{(t)})) = (\boldsymbol{\theta}^{(t-1)}, \mathbf{z}^{(t-1)})$,

7) Set $t = t + 1$,

8) If $t \leq N$ return to **3)**.

Quelques questions

- L'approximation ABC est une approximation non-paramétrique grossière de la densité *a posteriori*, peut-on faire mieux ?
- Comment choisir, l'ensemble des statistiques résumées, la distance, le seuil à coût numérique fixé...?
- Peut-on faire du ABC pour le choix de modèles ?

ABC et choix de modèles

Une collection M de modèles en compétition, pour $m \in \{1, \ldots, M\}$:
$f_m(\mathbf{y}|\boldsymbol{\theta}_m)$ and $\pi_m(\boldsymbol{\theta}_m)$
Une loi *a priori* sur les modèles: $\pi(\mathcal{M} = 1), \ldots, \pi(\mathcal{M} = M - 1)$

ABC algorithm for model choice

1) Set $i = 1$,

2) Generate m' from the prior $\pi(\mathcal{M} = m)$,

3) Generate $\boldsymbol{\theta}'_{m'}$ from the prior $\pi_{m'}(\cdot)$,

4) Generate $\mathbf{z}$ from the model $f_{m'}(\cdot|\boldsymbol{\theta}'_{m'})$,

5) If $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \epsilon$, set $m^i = m'$, $\boldsymbol{\theta}^i_{m^i} = \boldsymbol{\theta}'_{m'}$ and $i = i + 1$,

6) If $i \leq N$, return to **2)**.

Si $\eta(\mathbf{y})$ est une statistique exhaustive pour le choix de modèle [La loi conditionnelle de $\mathbf{y}$ sachant $\eta(\mathbf{y})$ est indépendante de l'indice du modèle] ABC-choix de modèle fonctionne parfaitement.

ABC likelihood-free methods for model choice in Gibbs random fields **Grelaud, Robert, Marin, Rodolphe and Taly (2009) Bayesian Analysis**

Asymptotiquement, contrairement à l'estimation des paramètres, la perte d'information (exhaustivité) peut détériorer le choix de modèles.

Lack of confidence in approximate Bayesian computation model choice **(Robert, Cornuet, Marin and Pillai (2011) PNAS)**

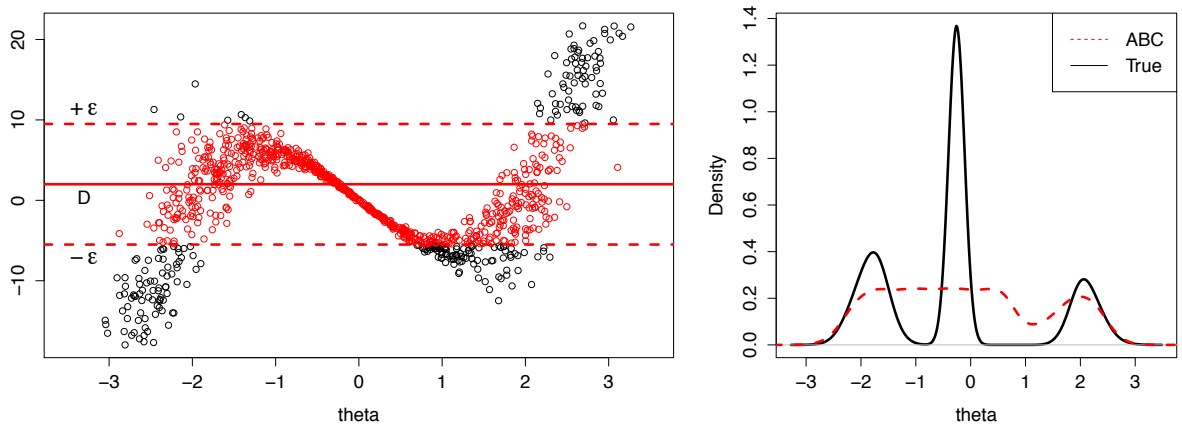
Exemples

Un cas d'école

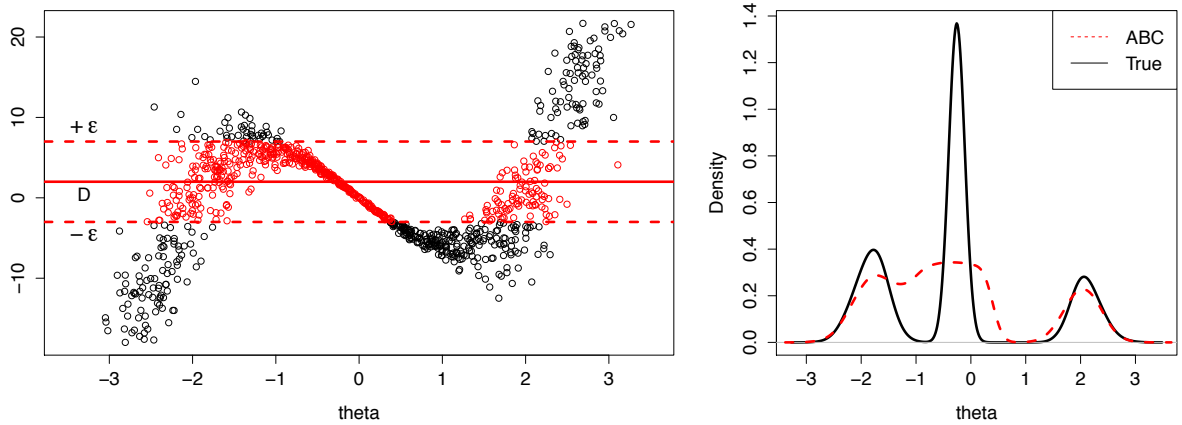
$y|\theta \sim \mathcal{N}_1\left(2(\theta+2)\theta(\theta-2), 0.1+\theta^2\right)$ and $\theta \sim \mathcal{U}_{[-10,10]}$

$y = 2$ $\rho(y,z) = |y-z|$

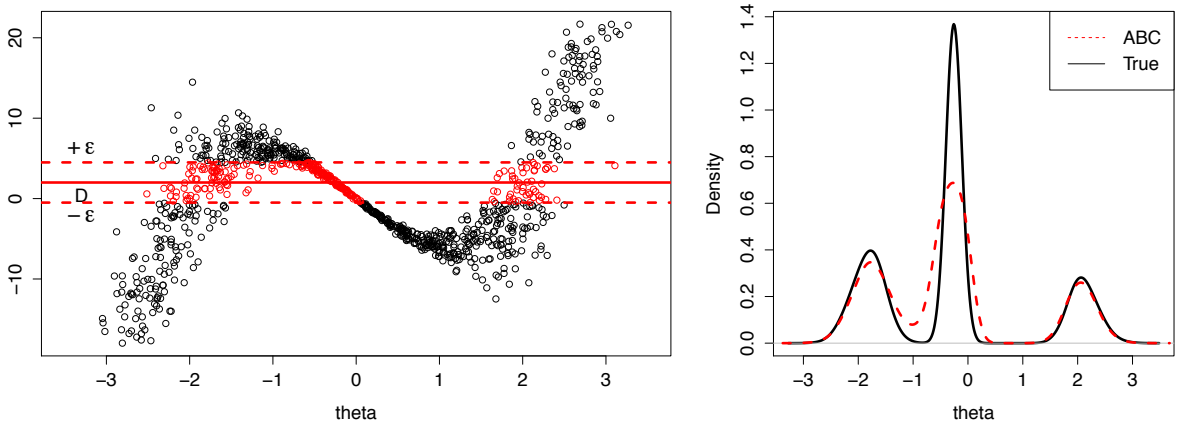
from Richard Wilkinson, Tutorial on ABC, NIPS 2013



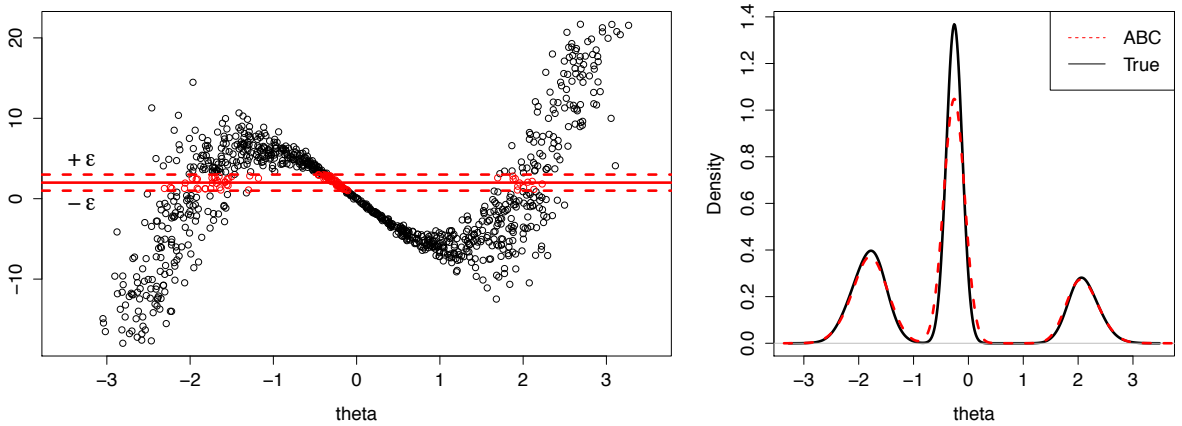
$\epsilon = 7.5$



$\epsilon = 5$

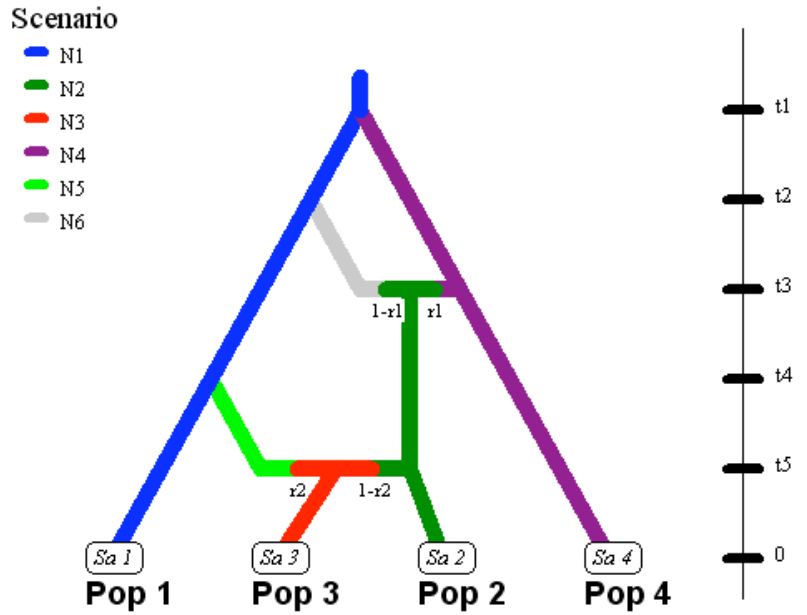


$\epsilon = 2.5$



$\epsilon = 1$

L'abeille européenne



Phylo-géographie de l'abeille européenne (*Apis mellifera*) depuis son aire d'origine (Asie orientale)

Deux voies d'invasion : l'une contournant les Alpes par le Nord et l'autre par le Sud : divergence à la date t_1, dont l'ordre de grandeur supposée est le demi million d'années.

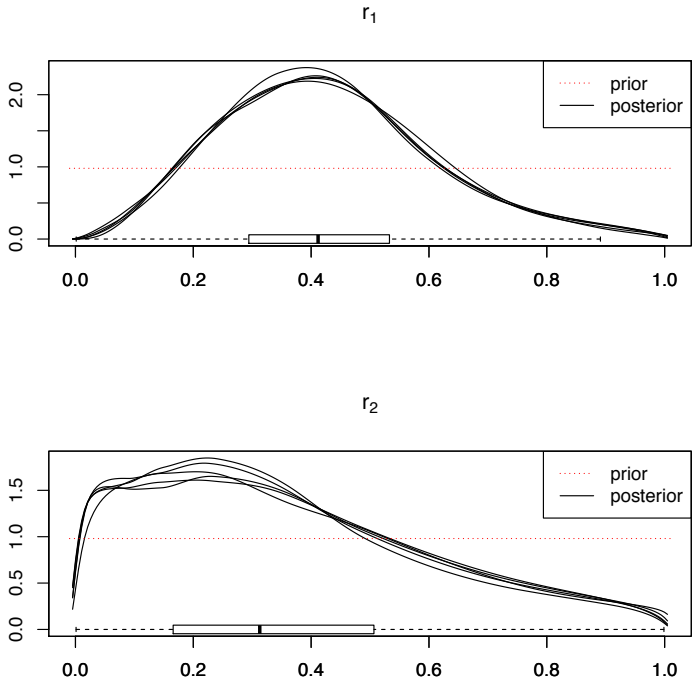
L'abeille présente en Italie (*Apis mellifera ligustica*) est un mélange entre la lignée *Apis mellifera mellifera* présente par exemple sur la côte occidentale française et la lignée *Apis mellifera carnica* que l'on retrouve en Europe du Sud Est : admixture à la date t_3.

Les trois sous-espèces apparaissent dans nos échantillons : Pop1 a été échantillonnée dans les Landes (France) pour représenter *Apis mellifera mellifera*, Pop2 a été échantillonnée en Lombardi (Italie) pour représenter *Apis mellifera ligustica* et Pop4 en Croatie pour représenter *Apis mellifera carnica*.

La dernière population échantillonnée provient des ruches de Courmayeur (Val d'Aoste). L'abeille résidente ici est plutôt *Apis mellifera*, mais a été enrichie récemment en gènes d'*Apis mellifera ligustica*, suite à l'introduction répétée de reines du centre de l'Italie par les apiculteurs : admixture à la date t_5.

Environ 50 individus par population

8 locus microsatellites indépendants



Taux d'admixtures

B.5 - Le logiciel DIYABC