Méthodes d'inférence de l'histoire démographique de populations structurées à partir de données de polymorphisme génétique

Jean-Michel Marin

Université Montpellier 2
Institut de Mathématiques et Modélisation de Montpellier (I3M)
Institut de Biologie Computationnelle (IBC)
Labex Numev

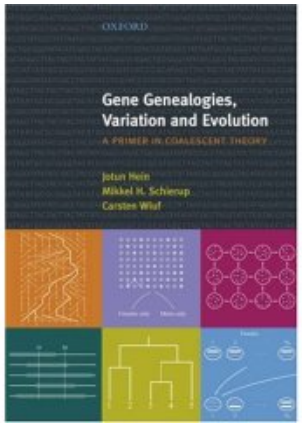
La génétique des populations

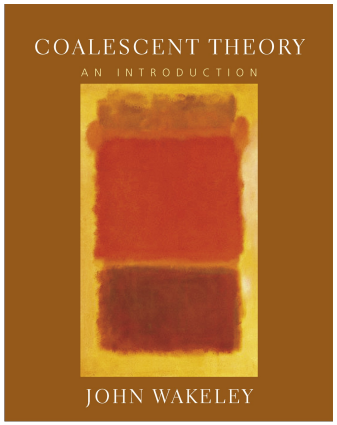
- Décrire les génotypes, estimer leur fréquence et celle des allèles, déterminer leur distribution au sein des individus, des populations, et entre les populations ;

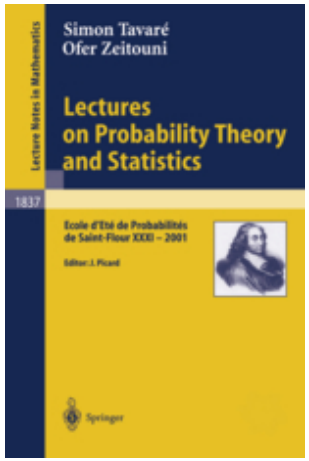
- Prédire et comprendre l'évolution des fréquences des gènes dans les populations sous l'effet de différents facteurs.

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

Un des développements principaux de la modélisation en génétique des populations est l'utilisation des méthodes dites coalescentes ou généalogiques.







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

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.

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

Il existe plusieurs options pour modéliser le lien entre ces différents locus : généalogie commune (liaison totale), généalogies partiellement partagées puis recombinaison ou généalogies indépendantes (aucune liaison).

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

La formulation de tels modèles est la première étape vers l'inférence statistique.
Ici, on s'intéresse à une classe de modèles probabilistes constitués d'événements inter-populationnels comme la divergence, l'admixture et la migration.

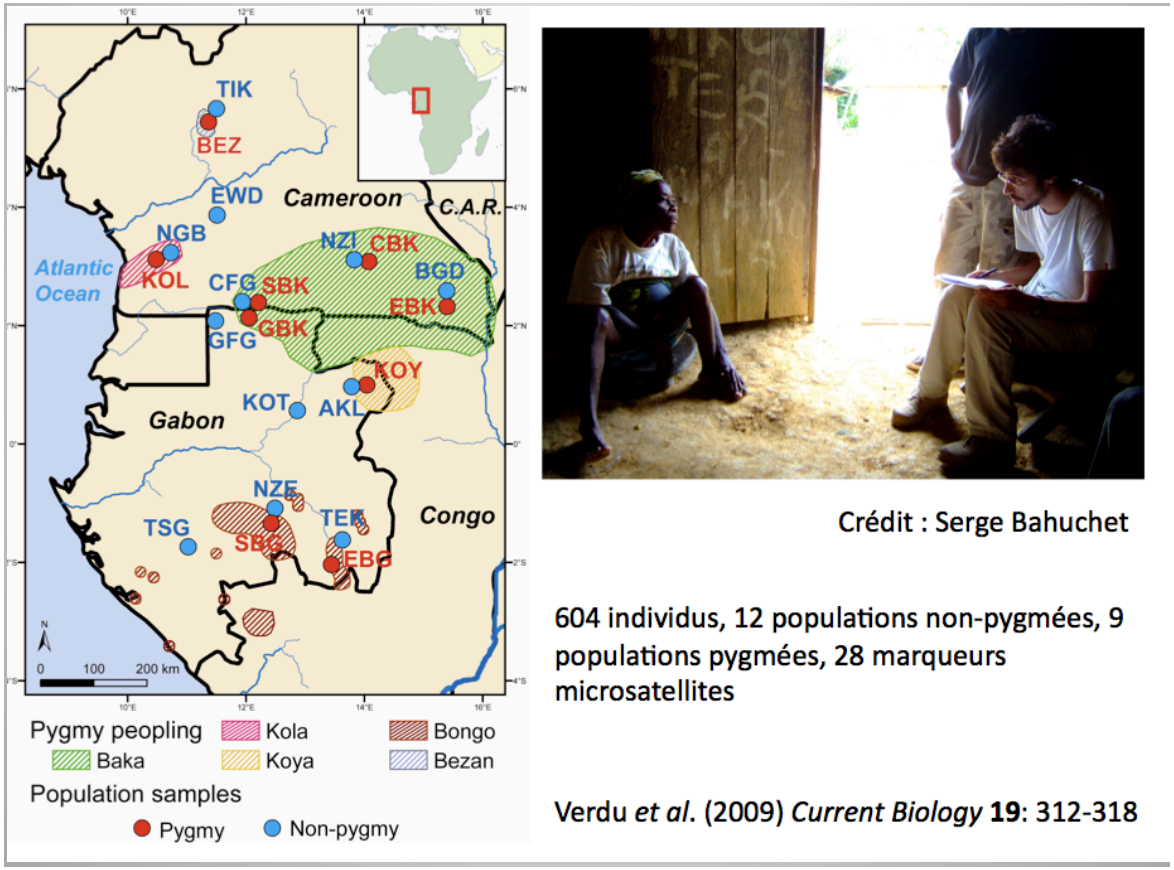
Les modèles que nous étudions sont sous l'hypothèse de neutralité Kimura (1968, 1983). Cette hypothèse implique l'absence d'effet de sélection.

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

Avec ces modèles, nous pouvons répondre à de nombreuses questions d'intérêt biologique : dater des divergences, quantifier des réductions ou des augmentations de tailles efficaces de populations, inférer des taux de migration, *etc*, **en mettant en place une procédure d'inférence des paramètres du modèle.**

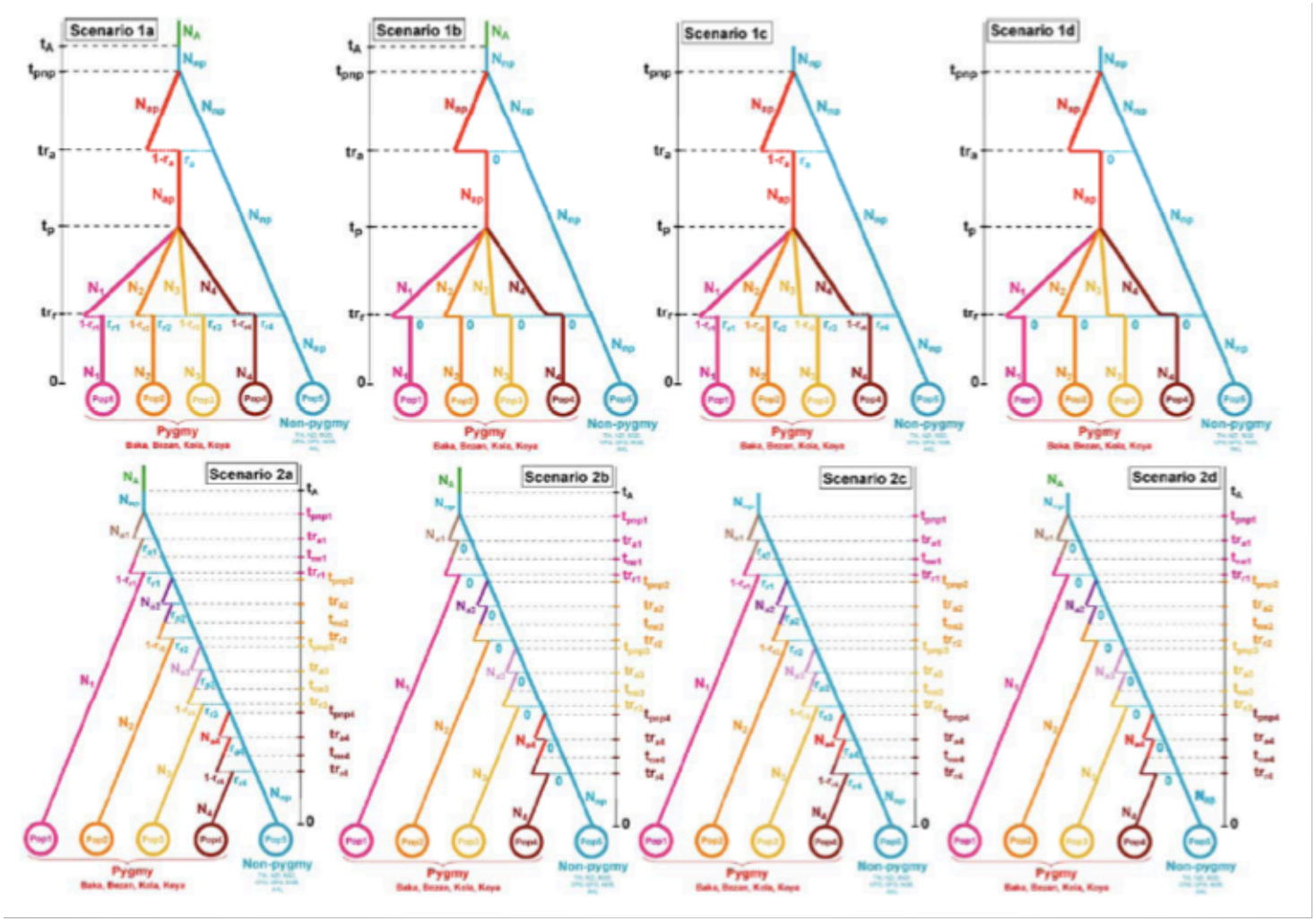
Nous pouvons également déterminer de quelle sources ancestrale provient une population récente, décrire des voies d'invasion de populations, *etc*. **Il faut alors utiliser une procédure de choix de modèle, chaque hypothèse correspond à un scénario démographique.**

Un exemple d'application sur les pygmées

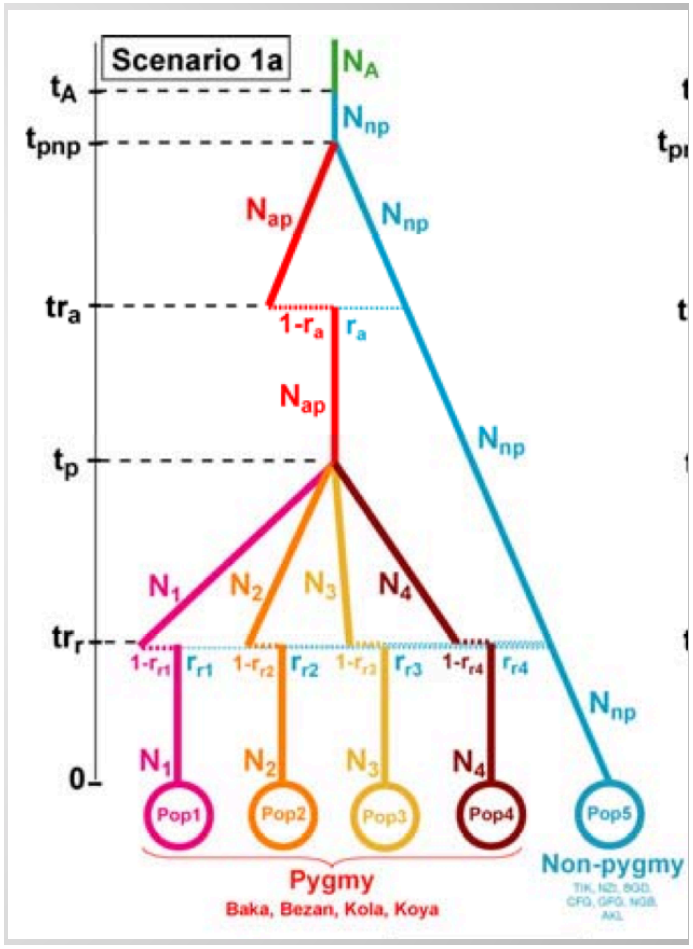


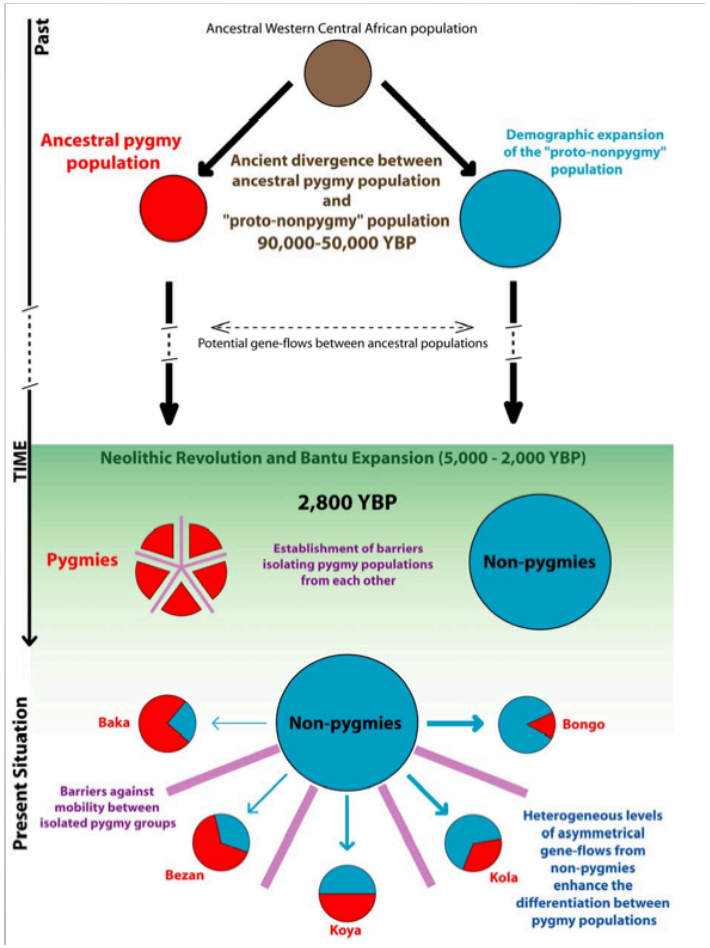
Les pygmées ont-ils une origine commune ? Y-a-t-il beaucoup d'échanges entre populations pygmées et non-pygmées ?

Différents scénarios possibles :



Le scénario 1a est choisi.





Un autre exemple d'application sur les orangs-outans

- 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 ?

Le séquençage du génome d'une femelle captive appelée Susie a été réalisé en janvier 2011. Il fait apparaître une similarité à 97% avec le génome humain.

Après les humains et les chimpanzés, les orangs-outans sont devenus la troisième espèce d'hominidés à avoir leur génome séquencé.

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.

Plan du cours

A - Modèles de génétique des populations (1h30)

A.1 - Données

A.2 - Généalogies d'échantillons

A.3 - Processus mutationnels

A.4 - Difficultés inférentielles

B - Méthodes bayésiennes approchées (3h30)

B.1 - Pré-requis de statistique bayésienne

B.2 - Méthodes ABC pour l'estimation de paramètres

B.3 - Méthodes ABC et choix de modèles

B.4 - Exemples

B.5 - Le logiciel DIYABC

A - Modèles de génétique des populations

A.1 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.

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**.

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**.

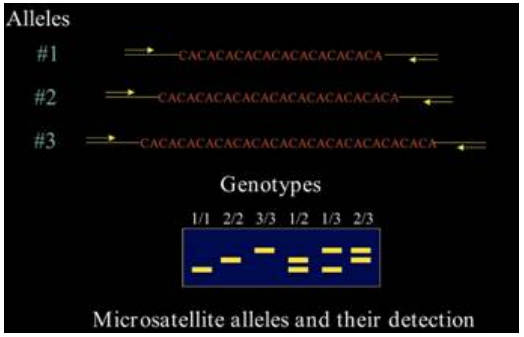
À ces locus, la séquence d'ADN peut varier d'un individu à l'autre, à cause des **mutations** au cours de l'évolution de l'espèce.

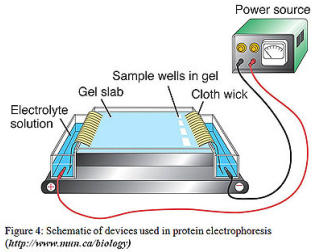
On parle alors de **polymorphisme génétique**. Les différentes variantes s'appellent des **allèles** ou des états alléliques.

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.

Marqueurs microsatellites : une partie de l'ADN où un court motif (de 1 à 4 paires de base) est répété en de nombreux exemplaires.



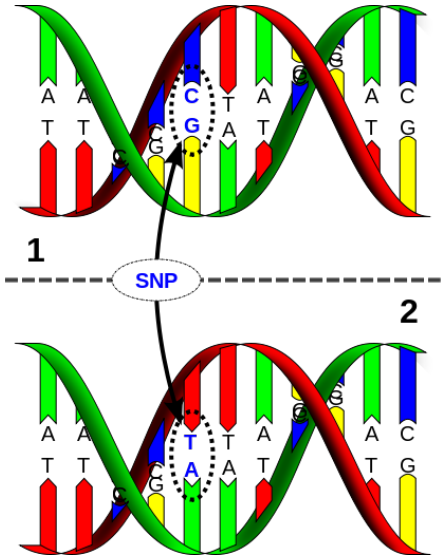


- **Microsatellite markers** :
  repetitions of short DNA motives,
  many loci dispersed in the genome
  - Medelian inheritance
  - high polymorphism
  - co-dominant
  - "neutral"

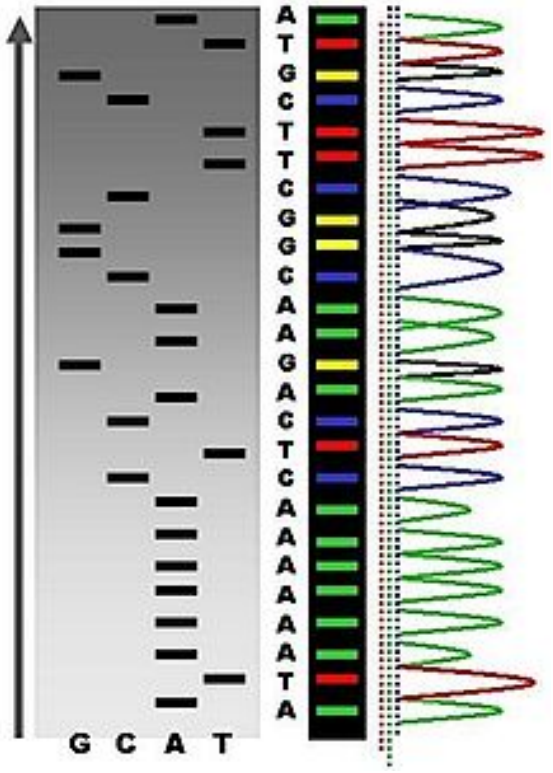
Fort polymorphisme, portion de l'ADN beaucoup utilisée en génétique des populations.

Marqueurs SNP : un type de polymorphisme de l'ADN dans lequel deux chromosomes diffèrent sur un segment donné par une seule paire de bases.

- **SNPs** : Single Nucleotide Polymorphisms
  - Medelian inheritance
  - low polymorphism
    0 / 1 = ancestral / derived states
  - many many SNPs in the genome
  - co-dominant
  - "neutral" or "selected" (good for the study of selection)



- **DNA sequences** : acces to the nucleotide sequence of "short" DNA fragments.
  - Medelian inheritance
  - intermediate polymorphism depending on the length
  - co-dominant but difficult to "phase"
  - "neutral" or "selected" : intra vs intergenic sequences



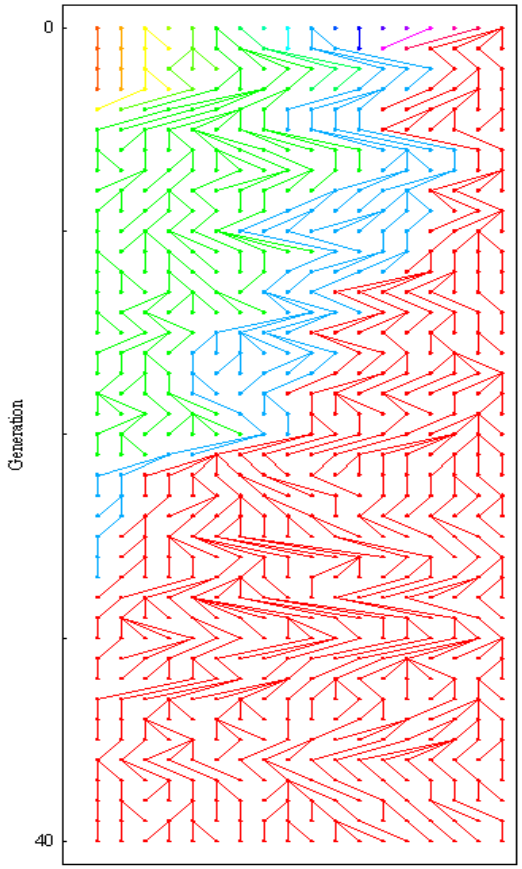
A.2 Généalogies d'échantillons

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.

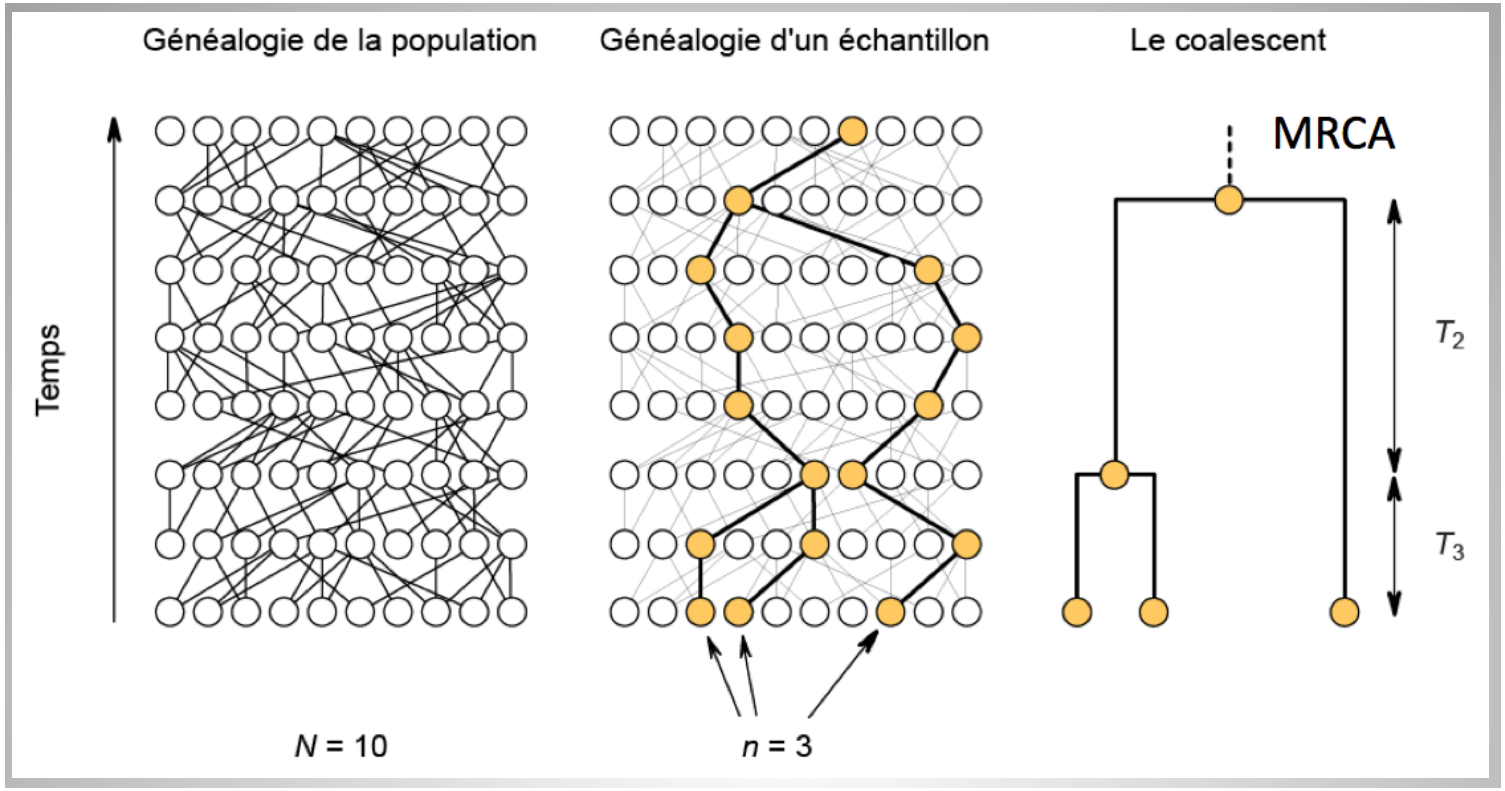
On s'intéresse seulement à l'évolution des ascendants des individus de notre échantillon en remontant le temps.

Le modèle de Wright-Fisher



- Une population de taille constante, dans laquelle les individus se reproduisent au même moment.
- Chaque gène à une génération est la copie d'un gène de la génération précédente.
- En l'absence de mutation et de sélection, les fréquences alléliques dérivent inévitablement jusqu'à la fixation d'un allèle.

Théorie de la coalescence (Kingman (1982), Tajima, Tavaré...)



La coalescence s'intéresse à la généalogie d'un échantillon de gènes en remontant le temps jusqu'à l'ancêtre commun de l'échantillon.

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.

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 $k\big(k-1\big)/2$.

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 $\frac{k\left(k-1\right)}{2Ne}$.

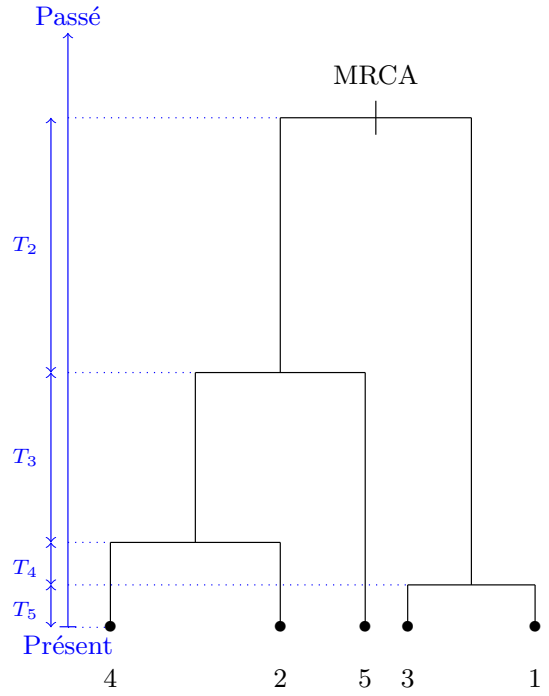
2) Augmenter les longueurs des k lignées de T_k.

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

4) $k \leftarrow k-1$.

Fin tantque

Cinq individus issus d'une seule population fermée à l'équilibre



Les individus échantillonnés sont représentés par les feuilles du dendrogramme, les durées inter-coalescences $T_2, \ldots, T_5$ sont indépendantes, et T_k est de loi exponentielle de paramètre $k\big(k-1\big)/2$.

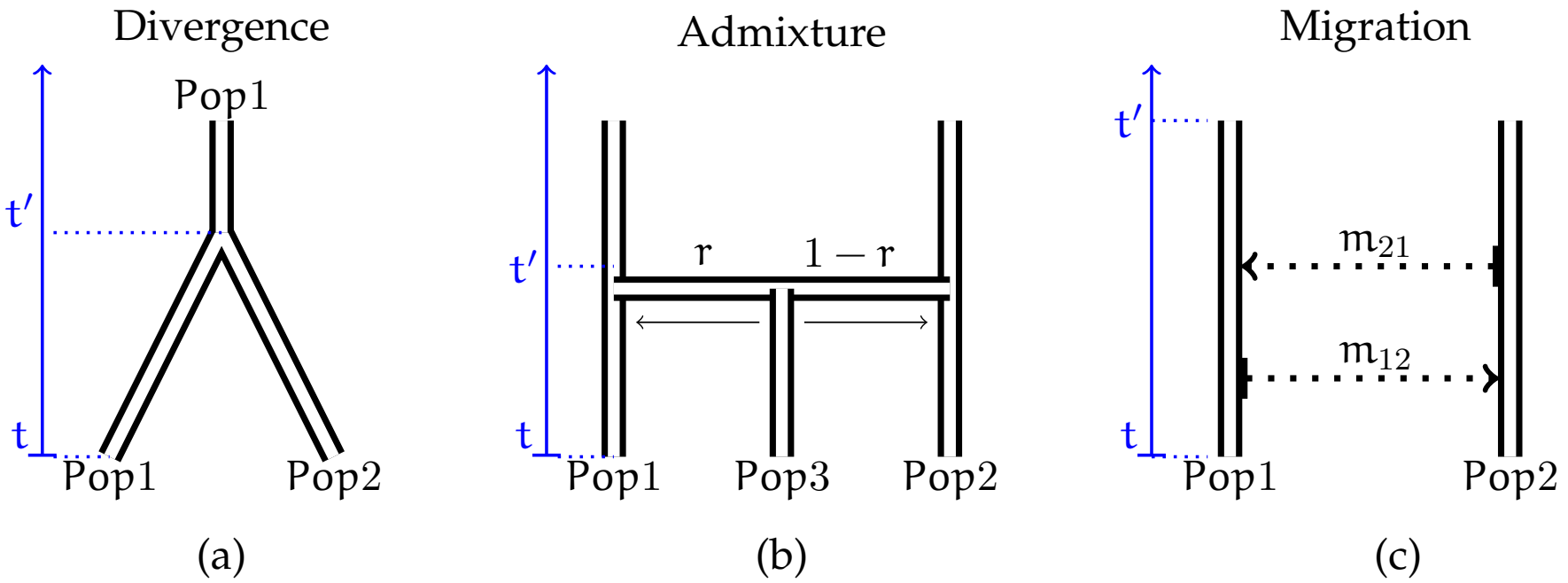
Plusieurs populations structurées

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.


Divergence
Pop1
t′
t
Pop1
Pop2
(a)
Admixture
t′
r
1 − r
t
Pop1
Pop3
Pop2
(b)
Migration
t′
m21
m12
t
Pop1
Pop2
(c)

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 $=1 \rightarrow D$ **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.

A.3 Processus mutationnels

Position des mutations

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.

Processus d'évolution à chaque mutation cas des microsatellites

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 $\mathbb{N}$.

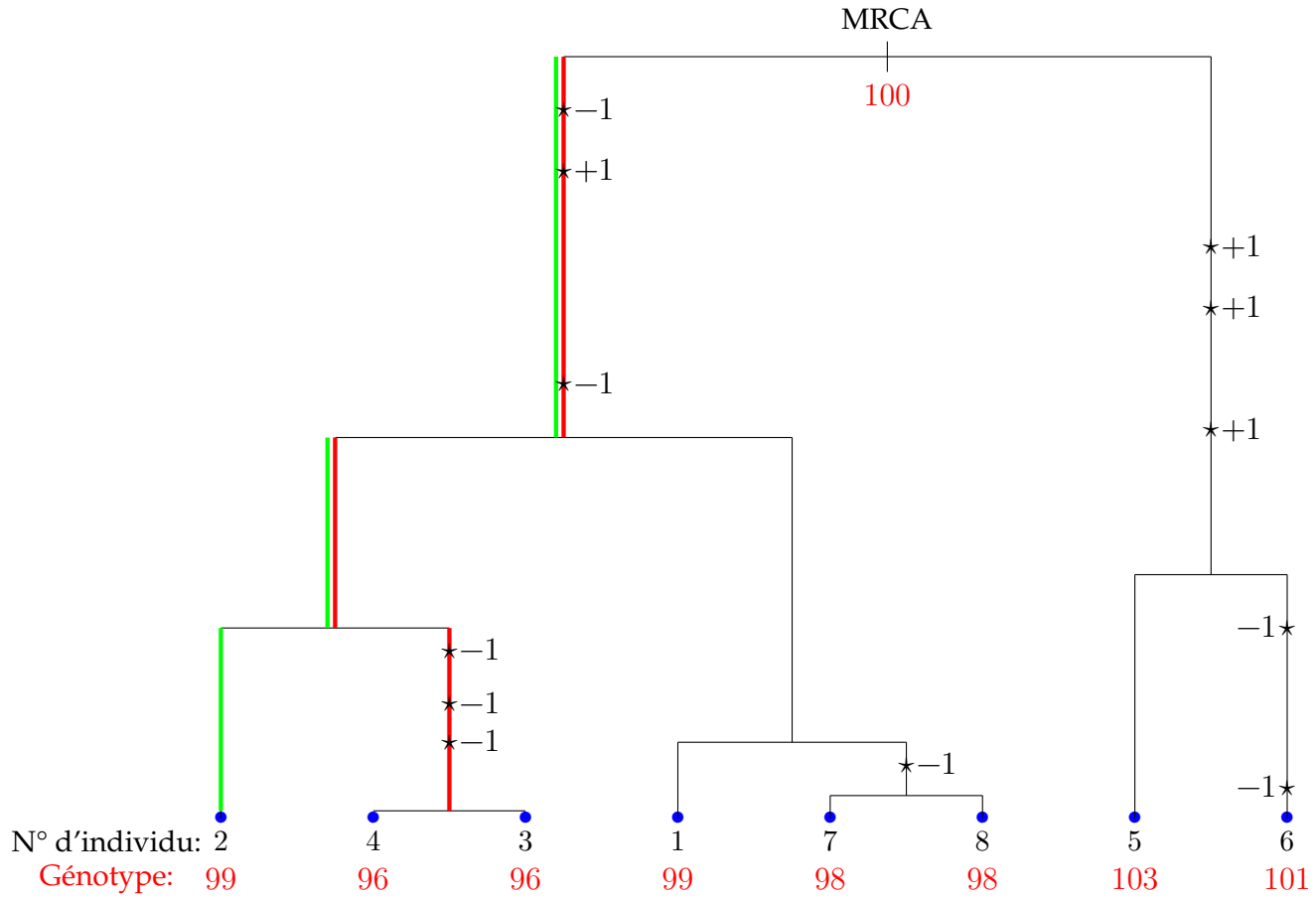
Modèle GSM : modifier le locus d'une longueur $\pm mG$, où m est la longueur connue du motif répété, G est une variable aléatoire de loi géométrique de paramètre p et un signe aléatoire (avec probabilités 1/2 et 1/2 respectivement).

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

A.4 Difficultés inférentielles

Chaque modèle est caractérisé par un ensemble de paramètres $\boldsymbol{\theta}$ historiques (temps de divergence, temps d'admixture, ...), démographiques (effectifs 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 :

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

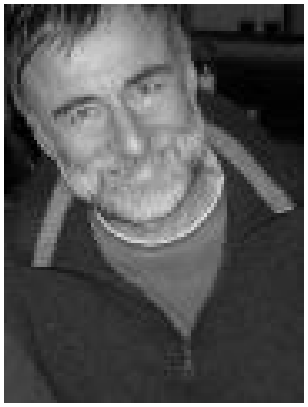
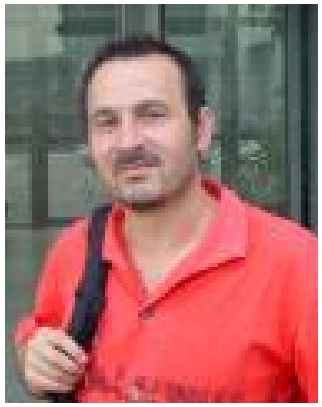
où $\mathbf{x}_i$ est l'ensemble des données au locus i et $\mathcal{M}_i \to \mathbf{x}_i$ désigne l'ensemble des génotypes sur le dendrogramme dont les feuilles correspondent à l'échantillon observé.

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.

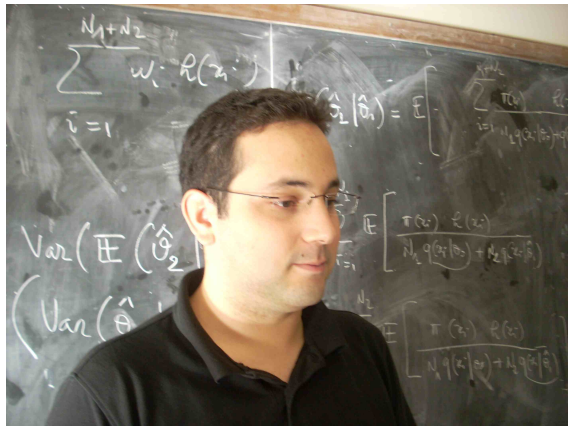
En dépit de la simplicité du coalescent de Kingman et du processus mutationnel, on ne peut espérer aucune simplification formelle dans cette intégrale.

- Coalescent Theory: An Introduction
  John Wakeley, 2008
- Gene Genealogies, Variation and Evolution:
  A primer in coalescent theory
  Jotun Hei, Mikkel Schierup, Carsten Wiuf, 2005
- Ancestral Inference in Molecular Biology
  Simon Tavaré, Saint-Flur, 2001

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



Coralie Merle, Pierre Pudlo, Christian Robert et Mohammed Sedki

B - Méthodes bayésiennes approchées

B.1 - 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 $\mathbf{x}$.

Dans le paradigme bayésien, le paramètre inconnu θ 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.

Le rôle central de la distribution a priori dans l'analyse statistique bayésienne ne réside pas dans le fait que le paramètre d'intérêt $\boldsymbol{\theta}$ puisse (ou ne puisse pas) être perçu comme étant distribué selon π, ou même comme étant une variable aléatoire,

mais plutôt dans la démonstration que l'utilisation d'une distribution a priori et de l'appareillage probabiliste qui l'accompagne est la manière la plus efficace [au sens de nombreux critères] de résumer l'information disponible (ou le manque d'information) sur ce paramètre ainsi que l'incertitude résiduelle.

Le seul moyen de construire une approche mathématiquement justifiée opérant conditionnellement aux observations est d'introduire une distribution correspondante pour les paramètres.

Le choix de la loi a priori est une difficulté majeure de l'approche bayésienne en ce que l'interprétation de l'information a priori disponible est rarement assez précise.

Signalons que la notion de loi a priori peut être étendue à des mesures de masse infinie tant que la loi a posteriori reste définie.

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 θ.

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

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

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

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.

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.

Méthodes de simulation

Let $(\mathbf{X}, \mathcal{B}(\mathbf{X}), \Pi)$ be a probability space.

(A1) $\Pi \ll \mu$ and $\Pi(dx) = \pi(x)\mu(dx)$.

(A2) π is known up to a normalizing constant:

- $\pi(x) = \dfrac{\tilde{\pi}(x)}{\displaystyle\int \tilde{\pi}(x)\mu(dx)}$;
- $\tilde{\pi}$ is known;
- the calculation of $\displaystyle\int \tilde{\pi}(x)\mu(dx) < \infty$ is intractable.

Problem: for some Π-measurable applications h, approximate

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

(A3) the calculation of $\int h(x)\tilde{\pi}(x)\mu(dx)$ is intractable.

Bayesian statistic: $\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Monte Carlo methods (MC)

$\Longrightarrow$ Generate an iid sample $x_1, \ldots, x_N$ from Π and to estimate $\Pi(h)$ by

$$\hat{\Pi}_N^{MC}(h) = N^{-1} \sum_{i=1}^{N} h(x_i).$$

$$\hat{\Pi}_N^{MC}(h) \longrightarrow_{as} \Pi(h)$$

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

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

Often impossible to simulate directly from Π!

Markov Chain Monte Carlo methods (MCMC)

$\Longrightarrow$ Generate $x^{(1)}, \ldots, x^{(T)}$ from a Markov chain $(x_t)_{t \in \mathbb{N}}$ with stationary distribution Π and estimate $\Pi(h)$ by

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

Convergence to the stationary distribution could be very slow!

Metropolis–Hastings algorithms

Metropolis–Hastings algorithms are generic (or down-the-shelf) MCMC algorithms, compared with the Gibbs sampler, in the sense that they can be tuned with a much wider range of possibilities.

Those algorithms are also a natural extension of standard simulation algorithms like accept-reject or sampling importance resampling in the sense that they are also based on a *proposal* distribution, a major difference being that this proposal is *Markovian,* with kernel density $q(x, y)$.

If the *target* distribution has density π, the generic Metropolis–Hastings algorithm is:

Initialization: Choose an arbitrary $x^{(0)}$

Iteration t:

1. Given $x^{(t-1)}$, generate $\tilde{x} \sim q(x^{(t-1)}, x)$
2. Calculate

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

3. With probability $\min(\rho(x^{(t-1)}, \tilde{x}), 1)$ accept $\tilde{x}$ and set $x^{(t)} = \tilde{x}$; otherwise reject $\tilde{x}$ and set $x^{(t)} = x^{(t-1)}$.

The distribution q is also called the *instrumental* distribution.

As in the Accept-Reject method, we only need to know both π and q up to proportionality, since the constant of proportionality of π cancel in the calculation of ρ.

Note also the advantage compared with the Gibbs sampler: it is not necessary to know the conditional distributions of π.

This algorithm only needs to simulate from q which we can choose arbitrarily, as long as q is capable of reaching all areas of positive probability under π.

While theoretical guarantees that the algorithm converges are very high, the choice of q remains paramount in practice.

Poor choices of q may indeed result either in a very high rejection rate, meaning that the Markov chain $(x^{(t)})_t$ hardly moves, or in a myopic exploration of the support of π, that is, in a dependence on the starting value $x^{(0)}$, with the chain stuck in a neighbourhood mode to $x^{(0)}$.

The random walk sampler

A *random walk* proposal has a symmetric transition density $q(x, y) = q_{RW}(y - x)$ where $q_{RW}(x) = q_{RW}(-x)$.

In this case the acceptance probability $\rho(x, y)$ reduces to the simpler form

$$\rho(x, y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right) .$$

The appeal of this scheme is obvious when looking at the acceptance probability, since it only depends on the target π and accepts all proposals that increase π.

There is a considerable flexibility in the choice of the distribution q_{RW}, at least in terms of scale (i.e., the size of the neighbourhood of the current value) and tails.

Example

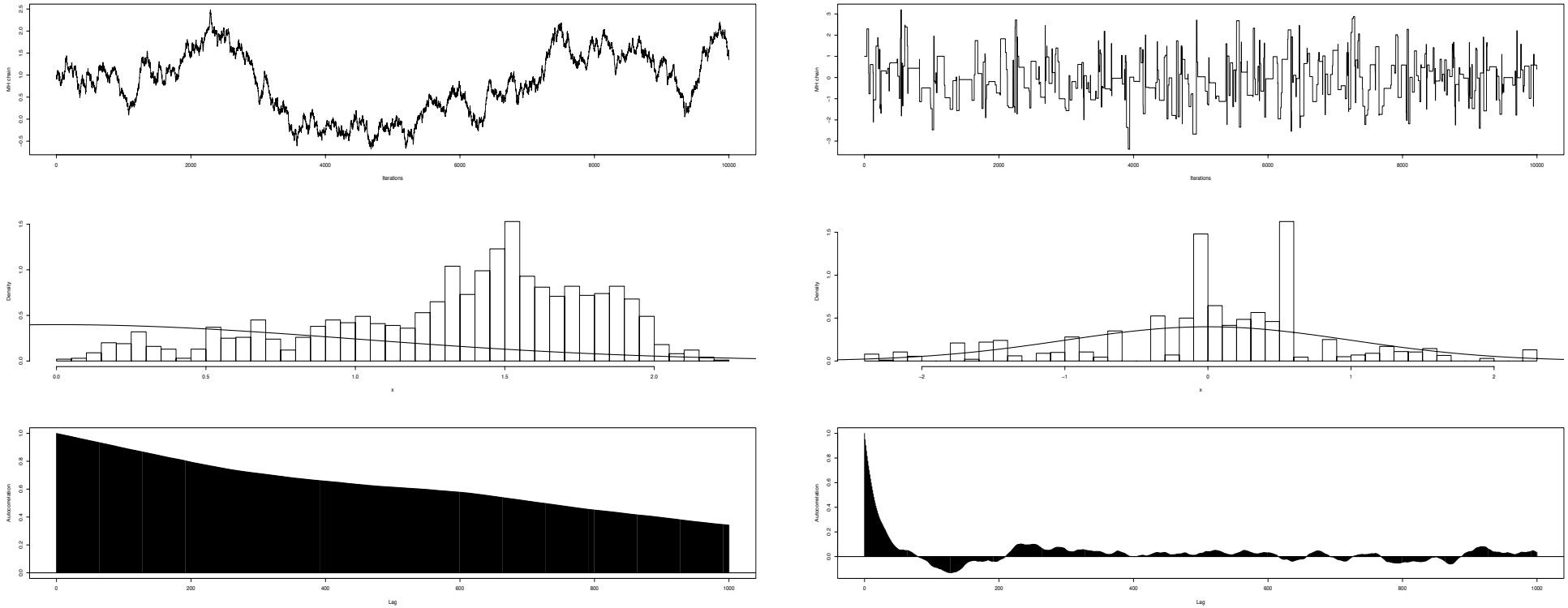
Consider the standard normal distribution $\mathcal{N}(0,1)$ *as a target.*

If we use random walk Metropolis-Hastings algorithm with a normal random walk, i.e.

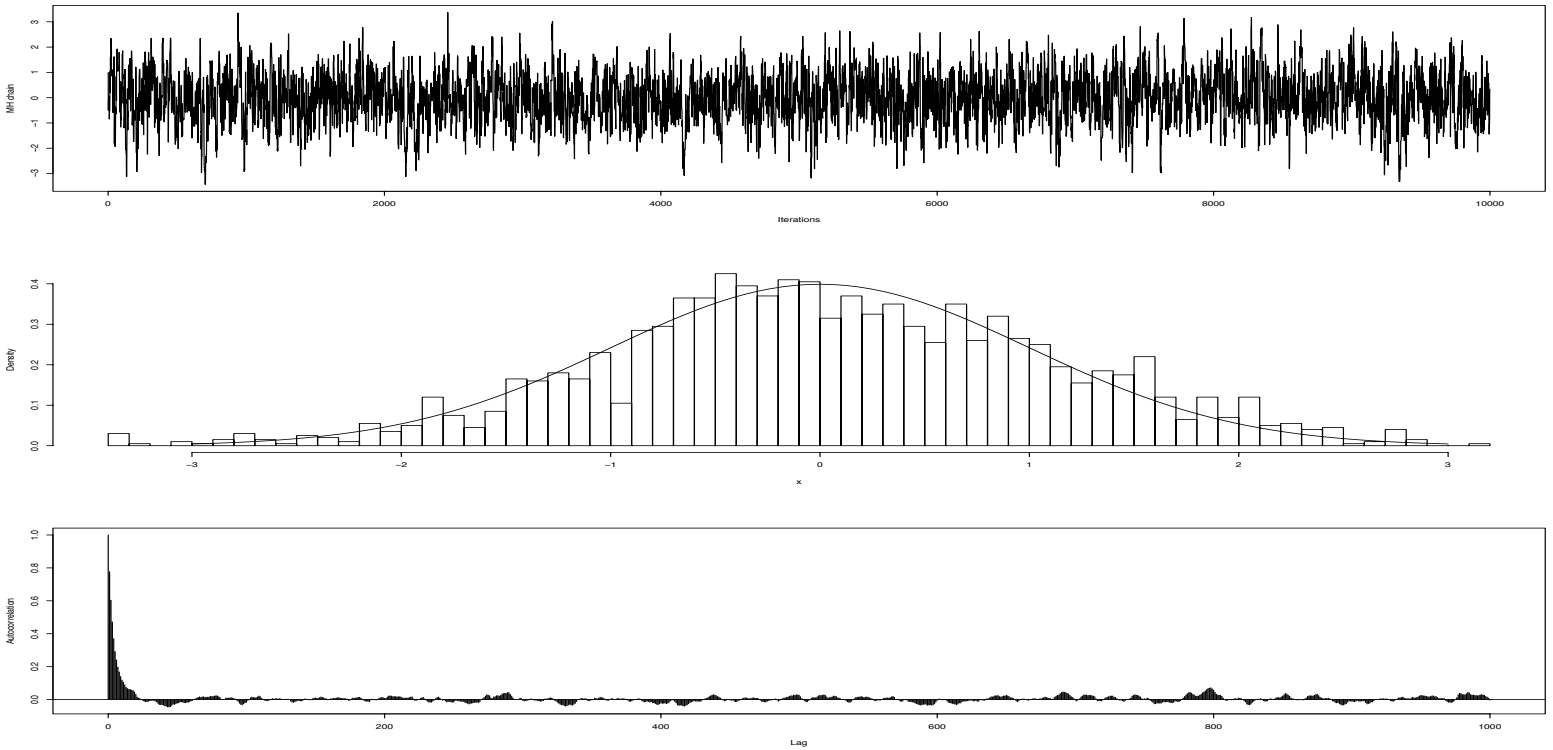
$$\tilde{x}|x^{(t-1)} \sim \mathcal{N}\left(x^{(t-1)}, \sigma^2\right) ,$$

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

the performances of the sampler depends on the value of σ^2*.*



(left) $\sigma^2 = 10^{-4}$ and a *(right)* $\sigma^2 = 10^3$ *Top*: sequence of $10,000$ iterations subsampled at every 10-th iteration; *middle*: histogram of the $2,000$ last iterations compared with the target density; *bottom*: empirical autocorrelations.


MH chain
Iterations
Density
x
Autocorrelation
Lag

Importance sampling

Let Q be a probability distribution on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$. Suppose that $\Pi \ll Q$, $Q \ll \mu$ and that $Q(dx) = q(x)\mu(dx)$:

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

$\Longrightarrow$ Generate an iid sample $x_1, \ldots, x_N$ from Q, called the proposal distribution, and to estimate $\Pi(h)$ by

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

$$\hat{\Pi}^{IS}_{Q,N}(h) \longrightarrow_{as} \Pi(h).$$

If $Q((h\pi/q)^2) < \infty$,

$$\sqrt{N}(\hat{\Pi}^{IS}_{Q,N}(h) - \Pi(h)) \longrightarrow_{\mathcal{L}} \mathcal{N}\left(0, Q((h\pi/q - \Pi(h))^2)\right).$$

For many h, a sufficient condition for $Q((h\pi/q)^2) < \infty$ is that π/q is bounded.

The normalizing constant of Π is unknown, not possible to use $\hat{\Pi}^{IS}_{Q,N}$.

It is natural to use the self-normalized version of the IS estimator,

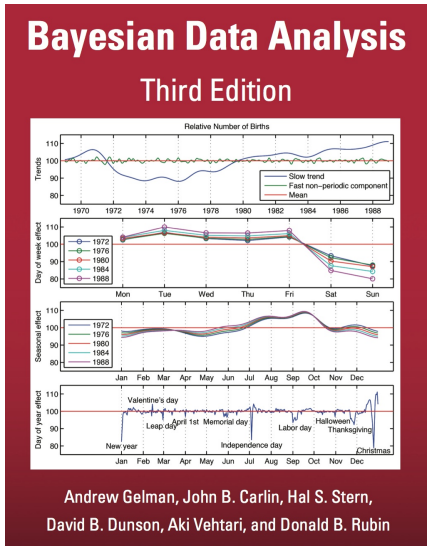
$$\hat{\Pi}^{SNIS}_{Q,N}(h) = \left(\sum_{i=1}^{N} \{\pi/q\}(x_i)\right)^{-1} \sum_{i=1}^{N} h(x_i)\{\pi/q\}(x_i).$$

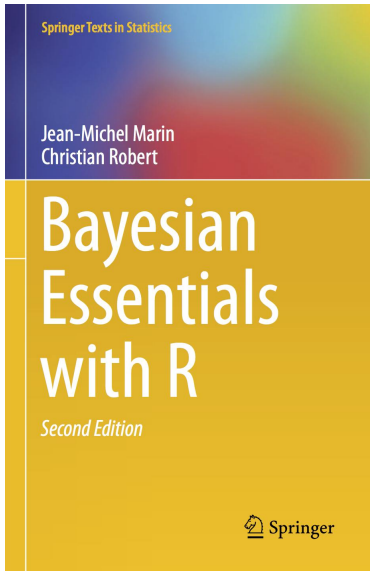
$$\hat{\Pi}_{Q,N}^{SNIS}(h) \longrightarrow_{as} \Pi(h).$$

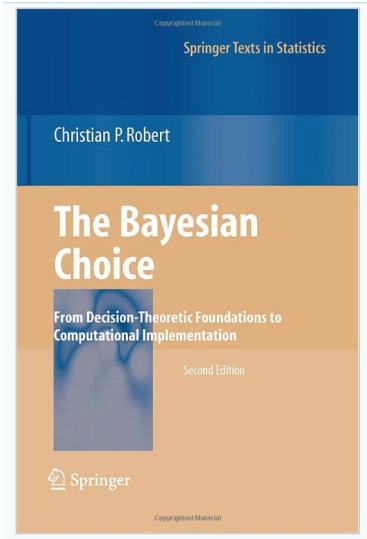
If $\Pi((1+h^2)(\pi/q)^2) < \infty$,

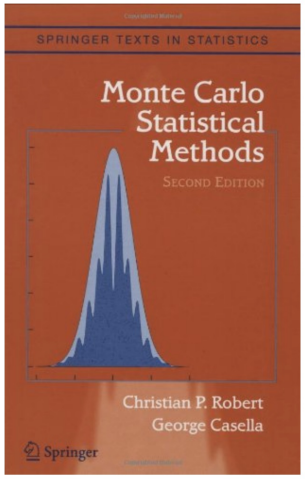
$$\sqrt{N}(\hat{\Pi}_{Q,N}^{SNIS}(h) - \Pi(h)) \longrightarrow_{\mathcal{L}} \mathcal{N}\left(0, Q((\pi/q)^2(h-\Pi(h))^2)\right).$$

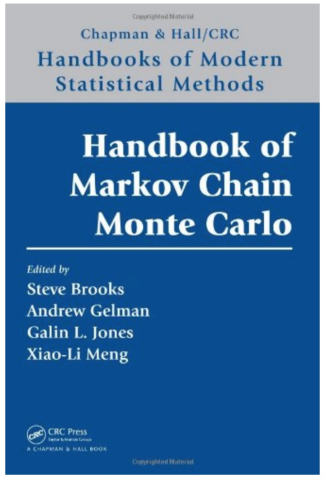
The quality of the SNIS approximation depends on the choice of the proposal distribution Q.

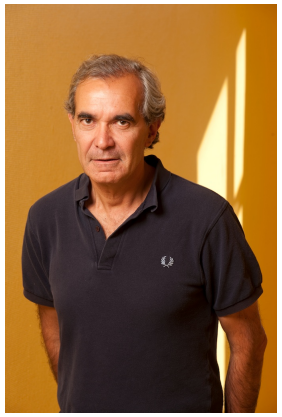

Bayesian Data Analysis
Third Edition
Relative Number of Births
Trends
Slow trend
Fast non−periodic component
Mean
Day of week effect
Seasonal effect
Day of year effect
Valentine's day
April 1st
Leap day
Memorial day
Labor day
Halloween
Thanksgiving
New year
Independence day
Christmas
Andrew Gelman, John 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

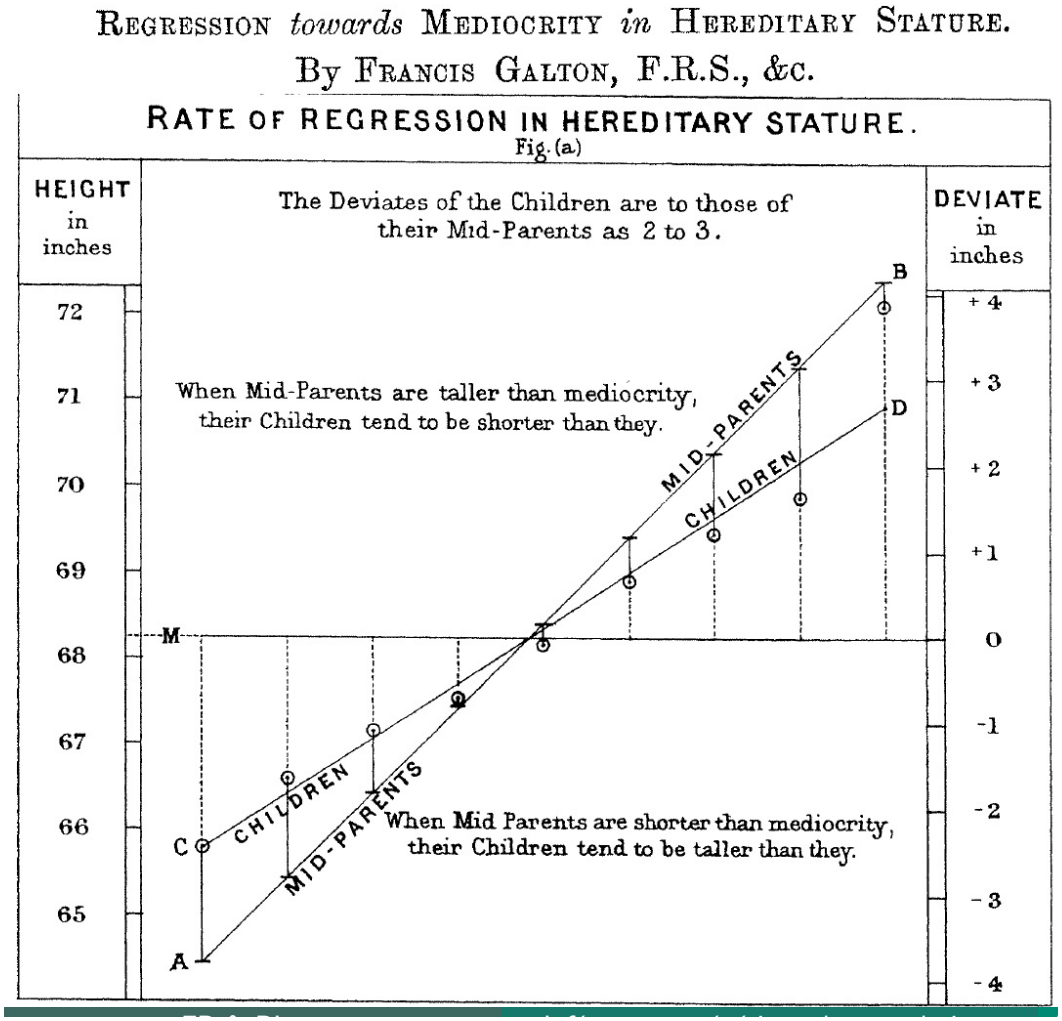

Chapman & Hall/CRC
Handbooks of Modern Statistical Methods
Handbook of Markov Chain Monte Carlo
Edited by
Steve Brooks
Andrew Gelman
Galin L. Jones
Xiao-Li Meng
CRC Press

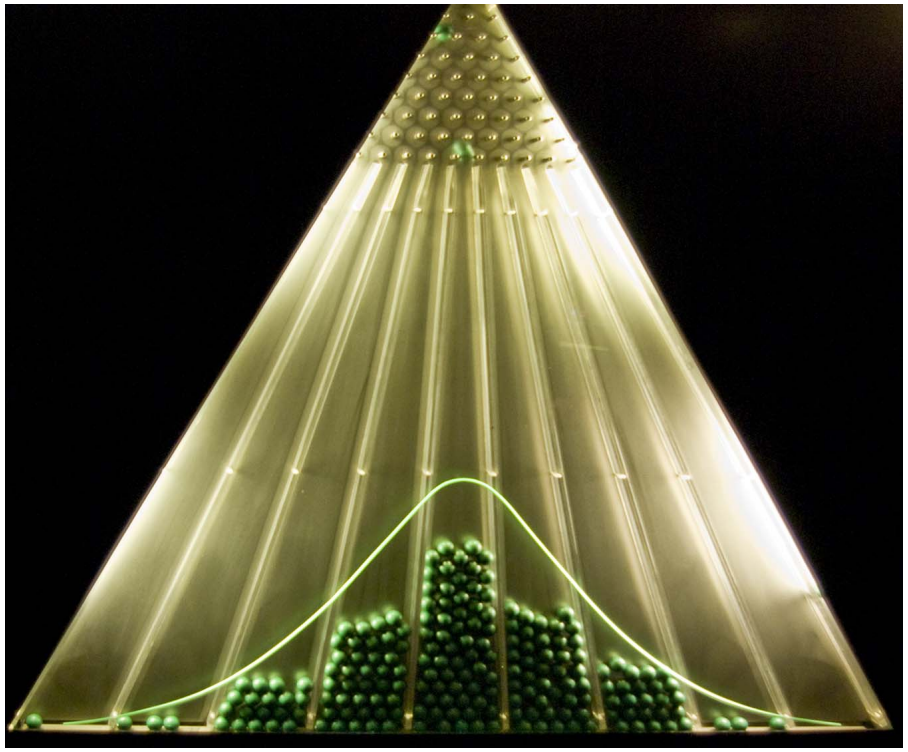



Olivier Cappé, Gilles Celeux, Randal Douc et Arnaud Guillin


Pierre Pudlo et Christian Robert

Francis Galton, cousin de Charles Darwin





Revue Biometrika...

B.2 - Méthodes ABC pour l'estimation de paramètres

Likelihood: $f(\mathbf{y}|\boldsymbol{\theta})$
Prior distribution on $\boldsymbol{\theta}$: $\pi(\boldsymbol{\theta})$

Suppose that $\mathbf{y}$ takes values in a countable set denoted by $\mathcal{D}$.

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

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

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

$(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N)$ is an iid sample from the posterior distribution.

The proof is trivial:

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

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

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

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)**.

The likelihood free algorithm sample from the marginal in $\mathbf{z}$ of:

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

- $\epsilon > 0$ a tolarance level,
- $\mathbb{I}_B(\cdot)$ the indicator function of a given set B,
- $A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \epsilon\}$.

The idea behind ABC is that the summary statistics coupled with a small tolerance should provide a good approximation of the posterior distribution:

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

A convolution approximation:

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

The ABC-MCMC method:

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)**.

The acceptance probability does not involve the calculation of the likelihood. Indeed,

$$\frac{\pi_\epsilon(\boldsymbol{\theta}', \mathbf{z}'|\mathbf{y})}{\pi_\epsilon(\boldsymbol{\theta}^{(t-1)}, \mathbf{z}^{(t-1)}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}') f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)}) f(\mathbf{z}'|\boldsymbol{\theta}')},$$

$$= \frac{\pi(\boldsymbol{\theta}') \cancel{f(\mathbf{z}'|\boldsymbol{\theta}')} \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}')}{\pi(\boldsymbol{\theta}^{(t-1)}) \cancel{f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})} \cancel{\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}^{(t-1)})}} \times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}') \cancel{f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)}) \cancel{f(\mathbf{z}'|\boldsymbol{\theta}')}},$$

$$= \frac{\pi(\boldsymbol{\theta}') q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t-1)} q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})} \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}').$$

For more details and method in the field of MCMC without likelihood, one can see the review paper of Sisson and Fan (2010) Handbook of Markov Chain Monte Carlo

Rejection sampling and MCMC methods can perform poorly if the tolerance level ϵ is small.

Consequently various sequential Monte Carlo algorithms have been constructed as an alternative to these two methods:
(Sisson et al. (2007) PNAS)
(Beaumont, Cornuet, Marin and Robert (2009) Biometrika)
(Del Moral et al. (2012) Statistics and Computing)
(Sedki, Cornuet, Marin, Pudlo and Robert (2014) under review)

The key idea is to decompose the difficult problem of sampling from $\pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ into a series of simpler subproblems.

The algorithm begins at time 0 sampling from $\pi_{\epsilon_0}(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ with large ϵ_0, then simulating from an increasing difficult sequence of target distribution $\pi_{\epsilon_t}(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$, that is $\epsilon_t < \epsilon_{t-1}$.

Regression adjustment

from Richard Wilkinson, Tutorial on ABC, NIPS 2013

An alternative to rejection-ABC, proposed by Beaumont *et al.* 2002, uses post-hoc adjustment of the parameter values to try to weaken the effect of the discrepancy between s and s_{obs}.

Two key ideas

- use non-parametric kernel density estimation to emphasise the best simulations
- learn a non-linear model for the conditional expectation $\mathbb{E}(\theta|s)$ as a function of s and use this to learn the posterior at s_{obs}.

Idea 1 kernel regression

Suppose we want to estimate

$$\mathbb{E}(\theta|s_{obs}) = \int \frac{\theta\pi(\theta, s_{obs})}{\pi(s_{obs})}\mathrm{d}\theta$$

using pairs $\{\theta_i, s_i\}$ from $\pi(\theta, s)$

Approximating the two densities using a kernel density estimate

$$\hat{\pi}(\theta, s) = \frac{1}{n}\sum_i K(s - s_i)K(\theta - \theta_i) \qquad \hat{\pi}(s) = \frac{1}{n}\sum_i K(s - s_i)$$

and substituting gives the Naradaya-Watson estimator:

$$\mathbb{E}(\theta|s_{obs}) \approx \frac{\sum_i K(s_{obs} - s_i)\theta_i}{\sum_i K(s_{obs} - s_i)}$$

as $\int yK(y - a)\mathrm{d}y = a$.

- Beaumont *et al.* 2002 suggested using the Epanechnikov kernel

$$K_\epsilon(x) = \frac{c}{\epsilon}\left[1 - \left(\frac{x}{\epsilon}\right)^2\right]\mathbb{I}_{x\leq\epsilon}$$

  as it has finite support - we discard the majority of simulations. They recommend ϵ be set by deciding on the proportion of simulations to keep e.g. best 5%
- This expression also arises if we view

$$\{\theta_i, W_i\}, \quad \text{with } W_i = K_\epsilon(s_{obs} - s_i) \equiv \pi_\epsilon(s_{obs}|s_i)$$

  as a weighted particle approximation to the posterior

$$\pi(\theta|s_{obs}) = \sum w_i \delta_{\theta_i}(\theta)$$

  where $w_i = W_i / \sum W_j$ are normalised weights
- The Naradaya-Watson estimator suffers from the curse of dimensionality - its rate of convergence drops rapidly as the dimension of s increases.

Idea 2 Regression adjustments

Consider the relationship between the conditional expectation of θ and s:

$$\mathbb{E}(\theta|s) = m(s)$$

Think of this as a model for the conditional density $\pi(\theta|s)$: for fixed s

$$\theta_i = m(s) + e_i$$

where $\theta_i \sim \pi(\theta|s)$ and e_i are zero-mean and uncorrelated

Suppose we've estimated $m(s)$ by $\widehat{m}(s)$ from samples $\{\theta_i, s_i\}$.
Estimate the posterior mean by

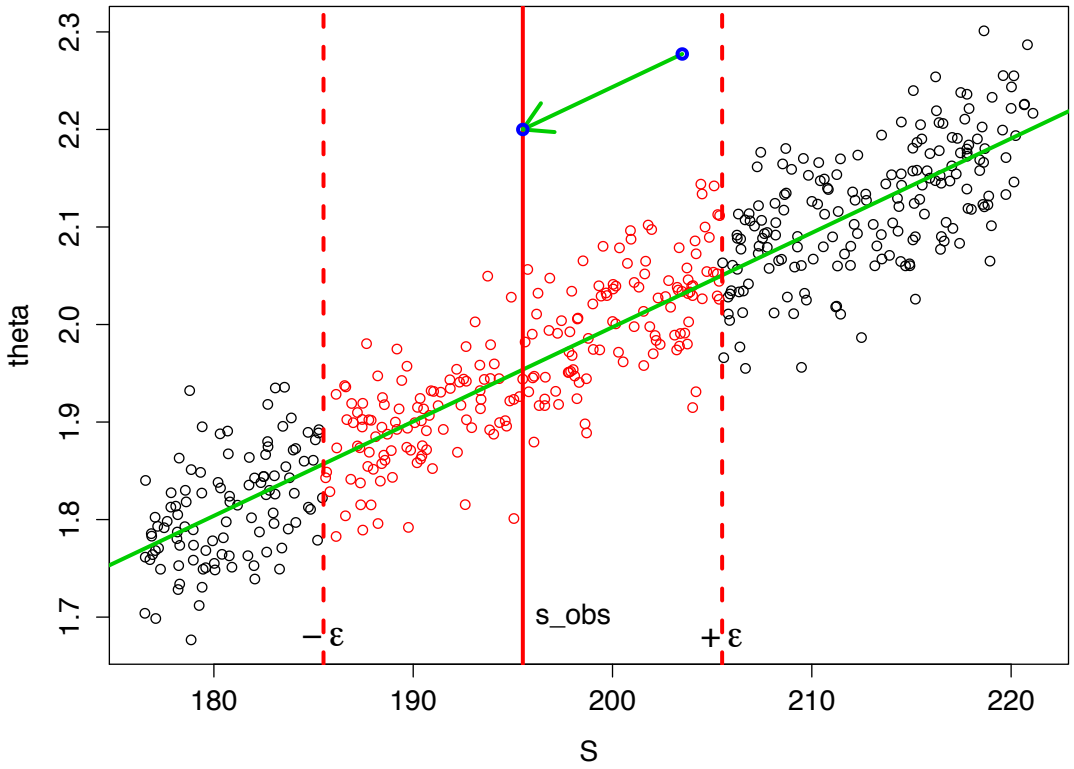
$$\mathbb{E}(\theta|s_{obs}) \approx \widehat{m}(s_{obs})$$

and assuming constant variance (wrt s), we can form the empirical residuals

$$\hat{e}_i = \theta_i - \widehat{m}(s_i)$$

and approximate the posterior $\pi(\theta|s_{obs})$ by adjusting the parameters

$$\theta_i^* = \widehat{m}(s_{obs}) + \hat{e}_i = \theta_i + (\widehat{m}(s_{obs}) - \widehat{m}(s_i))$$


theta
S
s_obs
−ε
+ε
1.7
1.8
1.9
2.0
2.1
2.2
2.3
180
190
200
210
220

Other questions

- The ABC rejection sampling scheme is associated to a crude non-parametric approximation of the posterior distribution, could we do better?
- How to choose the set of summary statistics, the distance, the tolerance level for a given computational effort...?
- ABC methods can handle model choice problems or not?

B.3 - Méthodes ABC et choix de modèles

Collection of M models, for $m \in \{1, \ldots, M\}$: $f_m(\mathbf{y}|\boldsymbol{\theta}_m)$ and $\pi_m(\boldsymbol{\theta}_m)$
Prior distribution for the models: $\mathbb{P}(\mathcal{M} = 1), \ldots, \mathbb{P}(\mathcal{M} = M - 1)$

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)**.

If $\eta(\mathbf{y})$ is a sufficient statistics for the model choice problem [the conditional distribution of $\mathbf{y}$ given $\eta(\mathbf{y})$ is independent of the model index] this can work pretty well.

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

Essentially no work on sufficiency and model choice.

Asymptotically speaking, contrary to the parameter estimation case, a loss of sufficiency can be a real difficulty for model choice questions.

Bayes factors based on summary statistics

Let us consider two models $f_1(\mathbf{y}|\boldsymbol{\theta}_1)$ and $f_2(\mathbf{y}|\boldsymbol{\theta}_2)$.

If $\eta_1(\mathbf{y})$ sufficient statistic for model $m = 1$ and parameter $\boldsymbol{\theta}_1$ and $\eta_2(\mathbf{y})$ sufficient statistic for model $m = 2$ and parameter $\boldsymbol{\theta}_2$, then $\eta(\mathbf{y}) = (\eta_1(\mathbf{y}), \eta_2(\mathbf{y}))$ is rarely sufficient for $(m, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

There is a fundamental discrepancy between the genuine Bayes factors/posterior probabilities and the Bayes factors based on summary statistics.

The genuine Bayes factor is the ratio of integrated likelihoods

$$B_{12}(\mathbf{y}) = \frac{\int \pi_1(\boldsymbol{\theta}_1) f_1(\mathbf{y}|\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}{\int \pi_2(\boldsymbol{\theta}_2) f_2(\mathbf{y}|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2} \,.$$

The Bayes factor based on $\eta(\mathbf{y})$ is the ratio of the integrated likelihoods for models defined on $\eta(\mathbf{y})$

$$B_{12}^{\eta}(\mathbf{y}) = \frac{\int \pi_1(\boldsymbol{\theta}_1) f_1^{\eta}(\eta(\mathbf{y})|\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}{\int \pi_2(\boldsymbol{\theta}_2) f_2^{\eta}(\eta(\mathbf{y})|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2} \,.$$

If $\eta(\mathbf{y})$ is a sufficient statistic for both models,

$$f_i(\mathbf{y}|\boldsymbol{\theta}_i) = g_i(\mathbf{y}) f_i^{\eta}(\eta(\mathbf{y})|\boldsymbol{\theta}_i)\,,$$

and

$$B_{12}(\mathbf{y}) = \frac{\int \pi_1(\boldsymbol{\theta}_1) g_1(\mathbf{y}) f_1^{\eta}(\eta(\mathbf{y})|\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}{\int \pi_2(\boldsymbol{\theta}_2) g_2(\mathbf{y}) f_2^{\eta}(\eta(\mathbf{y})|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2} = \frac{g_1(\mathbf{y})}{g_2(\mathbf{y})} B_{12}^{\eta}(\mathbf{y})\,.$$

Thus, unless $g_1(\mathbf{y}) = g_2(\mathbf{y})$ as in the special case of Gibbs random fields, the two Bayes factors differ by the ratio $g_1(\mathbf{y})/g_2(\mathbf{y})$.

The magnitude of the difference can be such that there is no direct connection between both answers.

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

$\mathcal{M}_1 : y \sim \mathcal{N}(\theta_1, 1)$;

$\mathcal{M}_2 : y \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$ (the Laplace or double exponential distribution with mean θ_2 and scale parameter $1/\sqrt{2}$, which has a variance equal to one).

$\mathbf{y} = (y_1, \ldots, y_n)$ is an iid sample from $\mathcal{M}_1$ or $\mathcal{M}_2$.

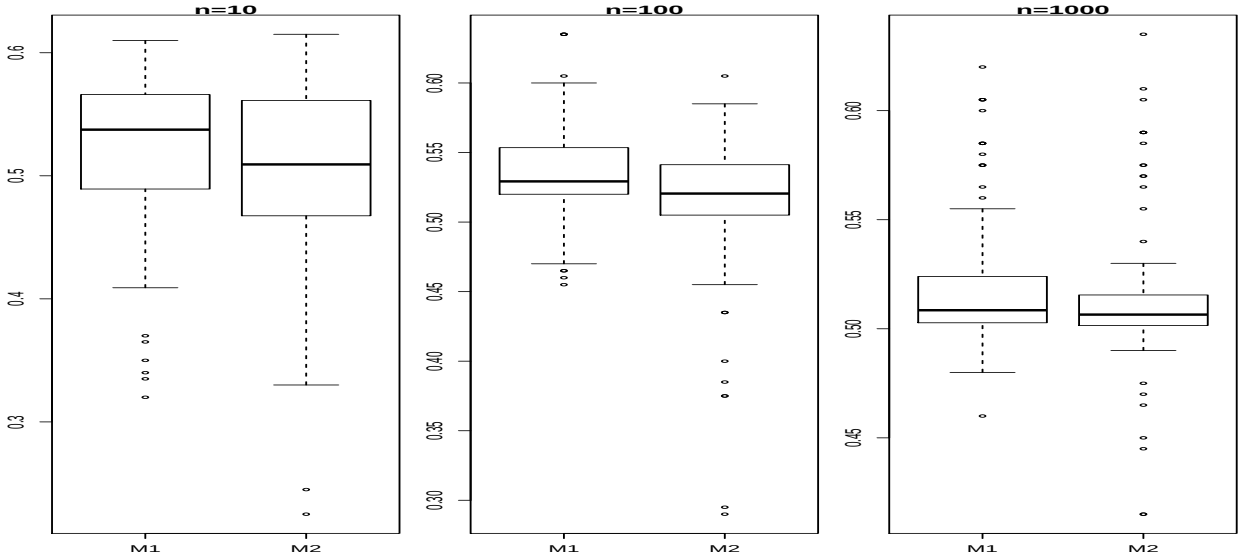
Four natural statistics can be considered:

1. the sample mean $\overline{\mathbf{y}}$: sufficient only for the Gaussian model;
2. the sample median $\text{med}(\mathbf{y})$: not sufficient but its distribution depends on θ_i in both models;
3. the sample variance $\text{var}(\mathbf{y})$: ancillary statistics;
4. the median absolute deviation $\text{mad}(\mathbf{y}) = \text{med}(|\mathbf{y} - \text{med}(\mathbf{y})|)$: ancillary statistics with different asymptotic expectations under $\mathcal{M}_1$ and $\mathcal{M}_2$;

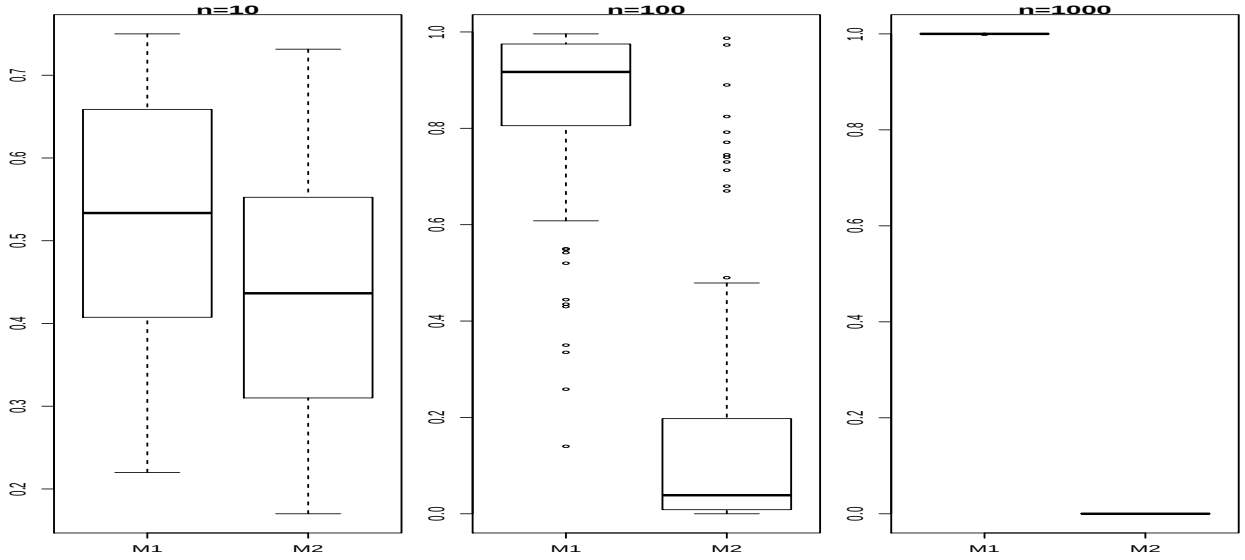
$\mathcal{M}_1$: Normal with $\theta_1 = 0$ and $\mathcal{M}_2$: Laplace with $\theta_2 = 0$

ABC algorithm: 10^4 proposals from the prior $\theta_i \sim \mathcal{N}(0, 4)$ and ϵ the 1% distance quantile

Summary statistic: the sample mean, median and variance



Summary statistic: the median absolute deviation



Relevant statistics for Bayesian model choice

(Marin, Pillai, Robert and Rousseau (2014) Journal of the Royal Statistical Society, Series B)

We derive sufficient conditions on summary statistics for the corresponding Bayes factor to be consistent, namely to asymptotically select the true model.




Mark Beaumont, Natesh Pillai, Christian Robert and Judith Rousseau

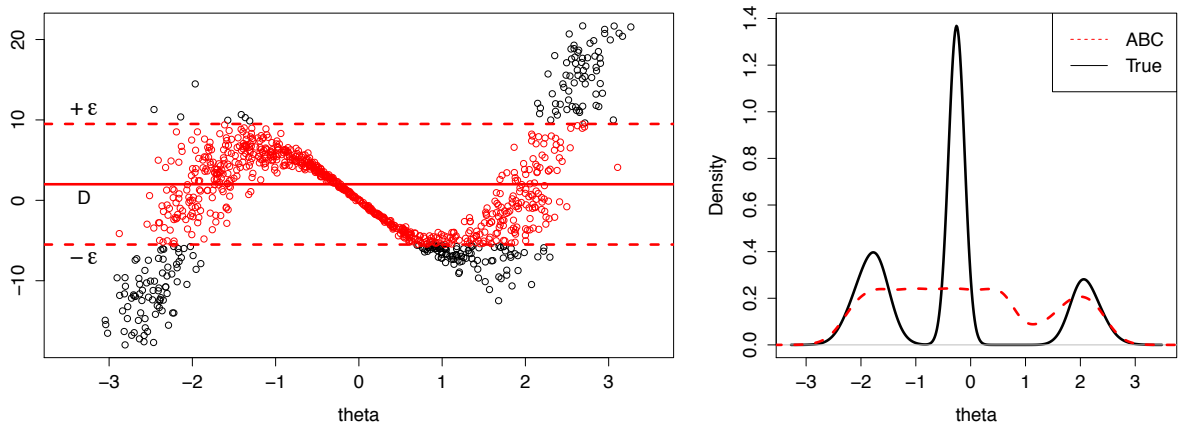
B.4 - Exemples

Un cas d'école

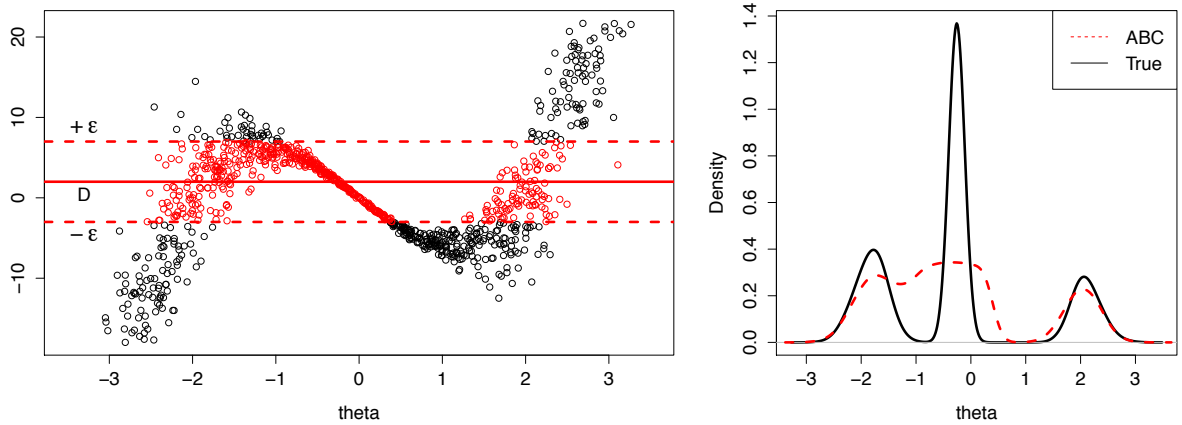
$y|\theta \sim \mathcal{N}_1\left(2(\theta+2)\theta(\theta-2), 0.1+\theta^2\right)$ and $\theta \sim \mathcal{U}_{[-10,10]}$

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

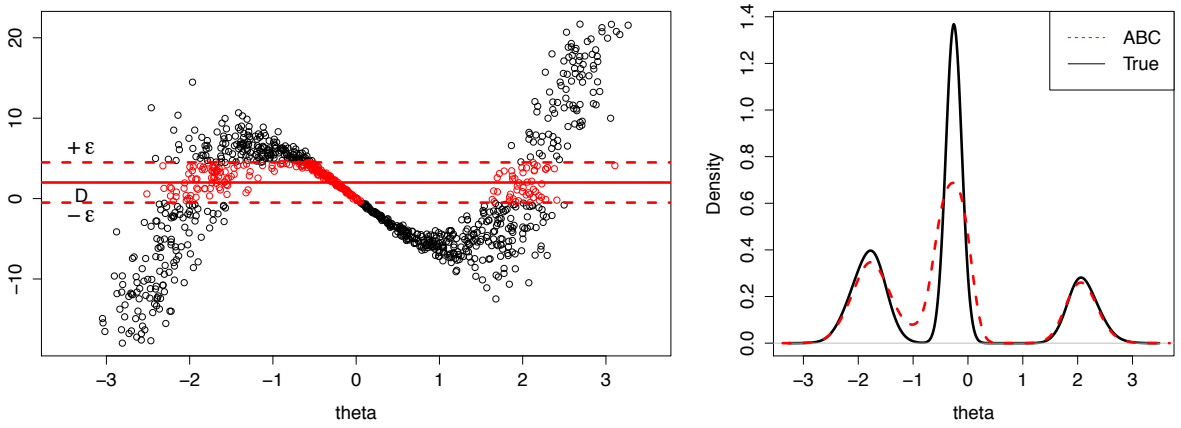
from Richard Wilkinson, Tutorial on ABC, NIPS 2013



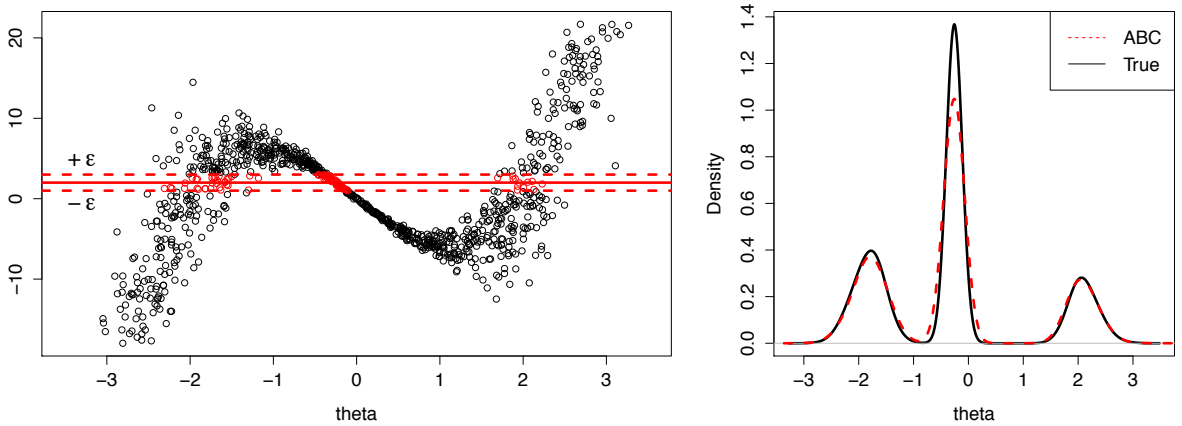
$\epsilon = 7.5$



$\epsilon = 5$

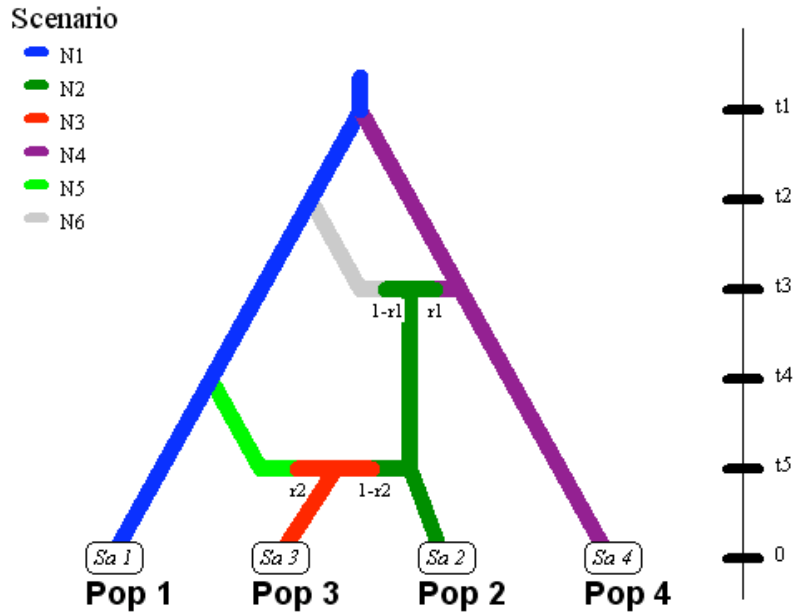


$\epsilon = 2.5$



$\epsilon = 1$

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.

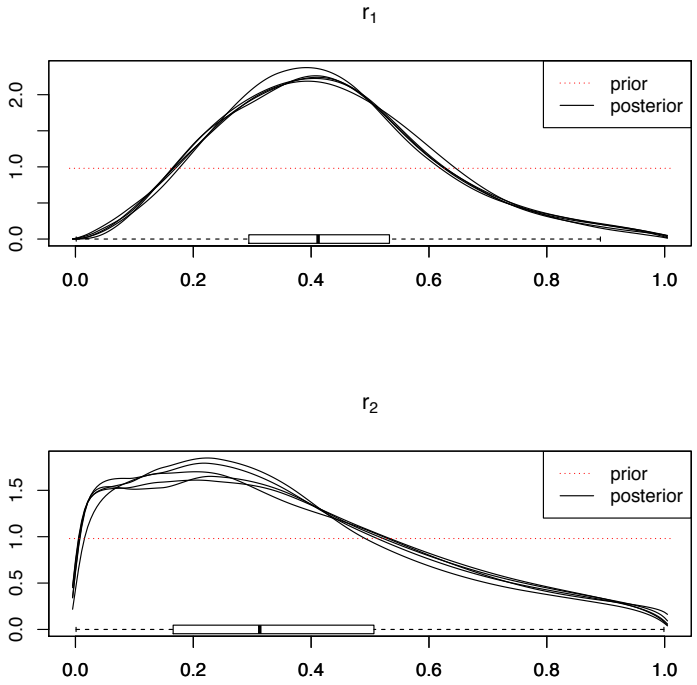
L'abeille présente en Italie (*Apis mellifera ligustica*) est un mélange entre la lignée *Apis mellifera mellifera* présente par exemple sur la côte occidentale française et la lignée *Apis mellifera carnica* que l'on retrouve en Europe du Sud Est : admixture à la date t_3.

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