Approximate Bayesian Computation methods for model choice a machine learning point of view

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck (IMAG)
Institut de Biologie Computationnelle (IBC)
Labex Numev

Juillet 2016

Introduction

When the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ is expensive or impossible to calculate, it is extremely difficult to sample from the posterior distribution

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

Two typical situations:

$f(\mathbf{y}|\boldsymbol{\theta}) = \displaystyle\int f(\mathbf{y}, \mathbf{u}|\boldsymbol{\theta}) \mu(\mathrm{d}\mathbf{u})$, the calculation of this integral is intractable and the latent vector $\mathbf{u}$ takes values in a high dimensional space (e.g. population genetics models).

$f(\mathbf{y}|\boldsymbol{\theta}) = g(\mathbf{y}, \boldsymbol{\theta})/Z(\boldsymbol{\theta})$ and the calculation of $Z(\boldsymbol{\theta})$ is intractable (e.g. for Markov random fields).

ABC is a technique that only requires being able to sample from the likelihood $f(\cdot|\boldsymbol{\theta})$.

This technique stemmed from population genetics models, about 15 years ago, and population geneticists still significantly contribute to methodological developments of ABC.

If, with Christian, we work on ABC methods, we can be very grateful to our biologist colleagues!

Likelihood-free rejection sampler

Rubin (1984) The Annals of Statistics

Tavaré et al. (1997) Genetics

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).
```

We keep the $\boldsymbol{\theta}$'s values such that the distance between the corresponding simulated dataset and the observed dataset is small enough.

The likelihood-free rejection sampler targets 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 tolerance level (threshold),
- $\mathbb{I}_B(\cdot)$ the indicator function of a given set B,
- $A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \epsilon\}$.

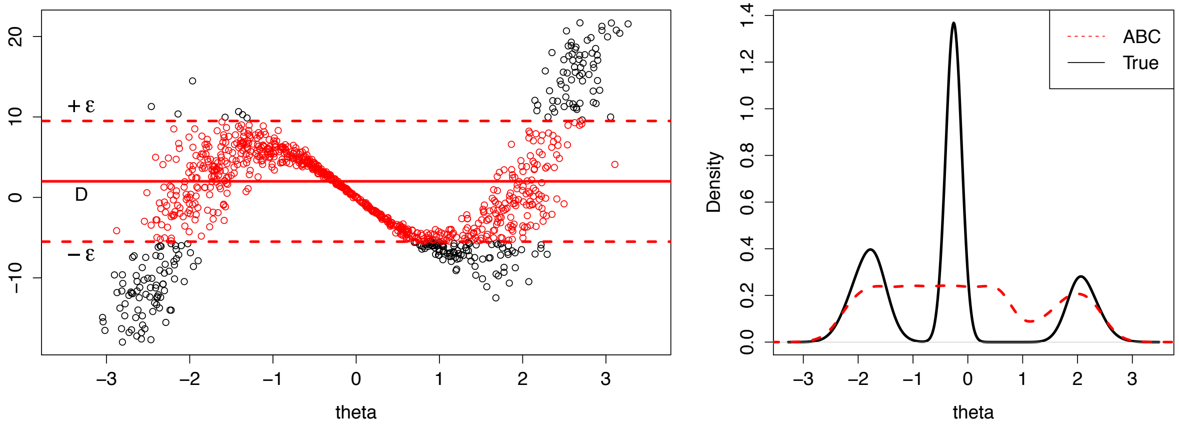
The idea behind ABC is that the summary statistics coupled with a small tolerance should provide a good approximation of the posterior distribution:

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

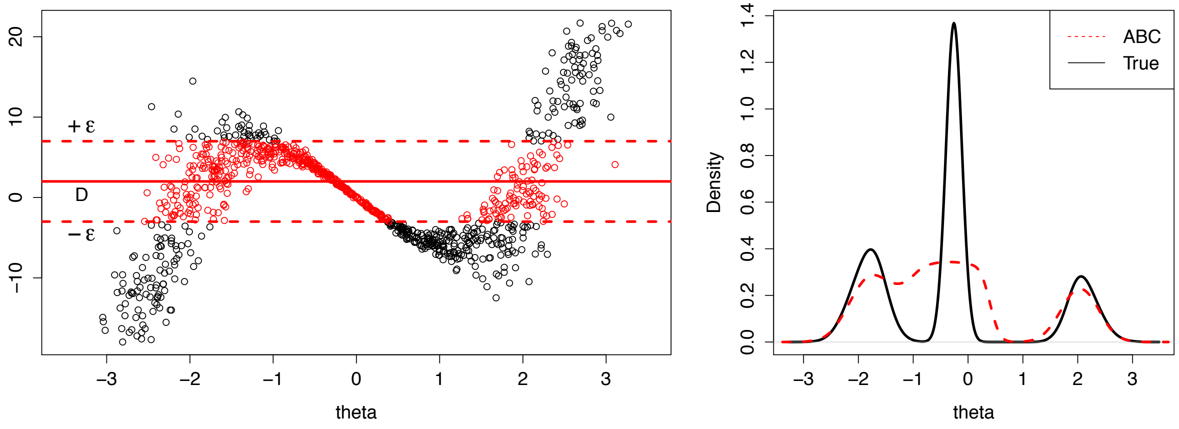
A toy example from Richard Wilkinson (Tutorial on ABC, NIPS 2013)

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

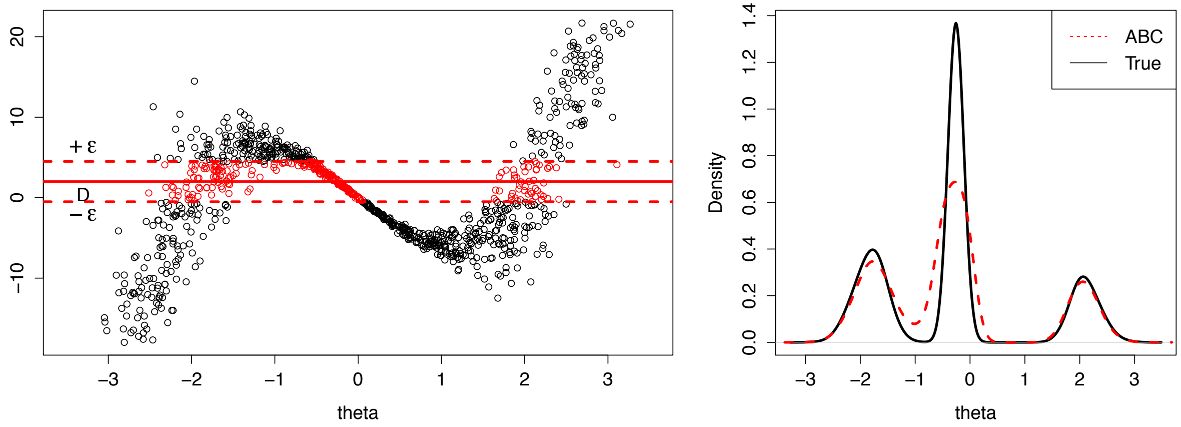
$y = 2 \ \rho(z, y) = |y - z|$



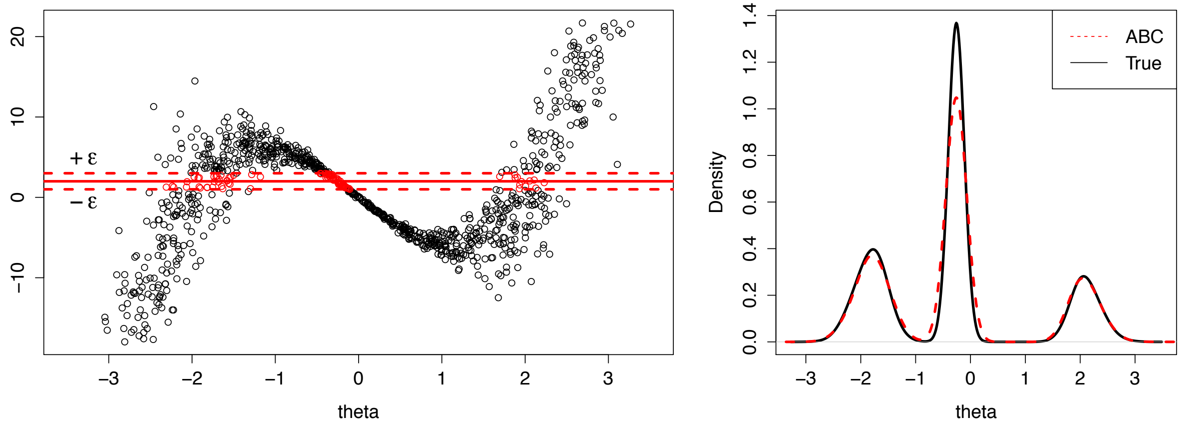
$\epsilon = 7.5$



$\epsilon = 5$



$$\epsilon = 2.5$$



$$\epsilon = 1$$

Practitioners really use $N = \lfloor \alpha M \rfloor$ and

```
1) For $i = 1, \dots, M$:
    a) Generate $\boldsymbol{\theta}_i$ from the prior $\pi(\cdot)$,
    b) Generate $\mathbf{z}$ from the model $f(\cdot|\boldsymbol{\theta}_i)$,
    c) Calculate $d_i = \rho(\eta(\mathbf{z}), \eta(\mathbf{y}))$,
2) Order the distances $d_{(1)}, \dots, d_{(M)}$,
3) Return the $\theta_i$'s that correspond to the $N$-smallest distances.
```

knn approximation, ϵ corresponds to a quantile of the distances.

- intuitive
- simple to implement
- embarrassingly parallelisable

Diggle and Gratton (1984) JRSS B suggested using a systematic simulation scheme to approximate the likelihood function.

They used a grid in the parameter space, several simulations for each grid point and apply smoothing techniques.

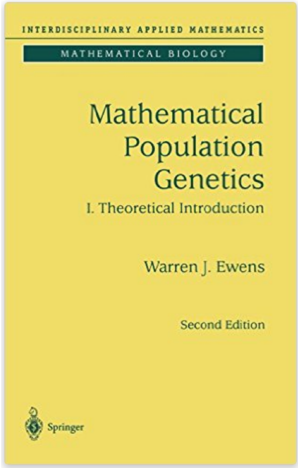
The term Approximate Bayesian Computation was established by **Beaumont et al. (2002) Genetics** extending further the ABC methodology and discussing the suitability of the ABC-approach more specifically for problems in population genetics.

Beaumont (2003) Genetics introduces pseudo-marginal MCMC algorithms which are approximations to an idealized marginal algorithm which can share the same marginal stationary distribution as the idealized method.

Population genetics motivations

Population genetics is concerned with the causes and effects of genetic variation.

One of the main developments in population genetics modeling is the use of coalescent methods.



INTERDISCIPLINARY APPLIED MATHEMATICS

MATHEMATICAL BIOLOGY

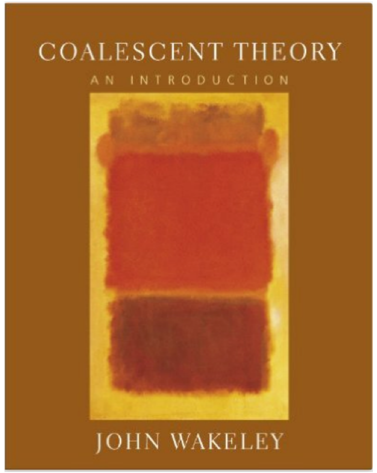
Mathematical Population Genetics

I. Theoretical Introduction

Warren J. Ewens

Second Edition

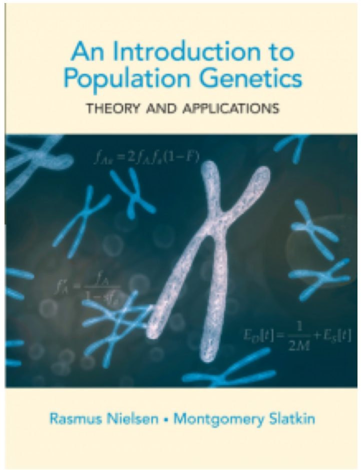
Springer



COALESCENT THEORY

AN INTRODUCTION

JOHN WAKELEY



An Introduction to Population Genetics

THEORY AND APPLICATIONS

Rasmus Nielsen • Montgomery Slatkin

The goal is to recover some elements of populations history. To analyse the structure of genetic data, these methods use the gene trees.

The formulation of a model is constrained by an evolutionary scenario that mimics the historical and demographic reality.

Such a scenario summarizes the evolutionary history of populations by a sequence of demographic events from an ancestral population.

Our datasets are composed of genetic informations coming from several locus, more and more locus...

There are several options to model the relationship between these different loci: common genealogy, partially shared genealogies and recombination, or independent genealogies.

For neutral models (**Kimura (1968, 1983)**, there is no selection effect. The observed polymorphisms are the result of genetic mutations on the genealogy of individuals.

With these models, we can answer questions of biological interest:

- estimate divergence times, quantify reductions or increases in effective population sizes, infer migration rates...
  parameter estimation problems

- determine from which ancestral sources comes a population, describe the invasion routes...
  model choice questions

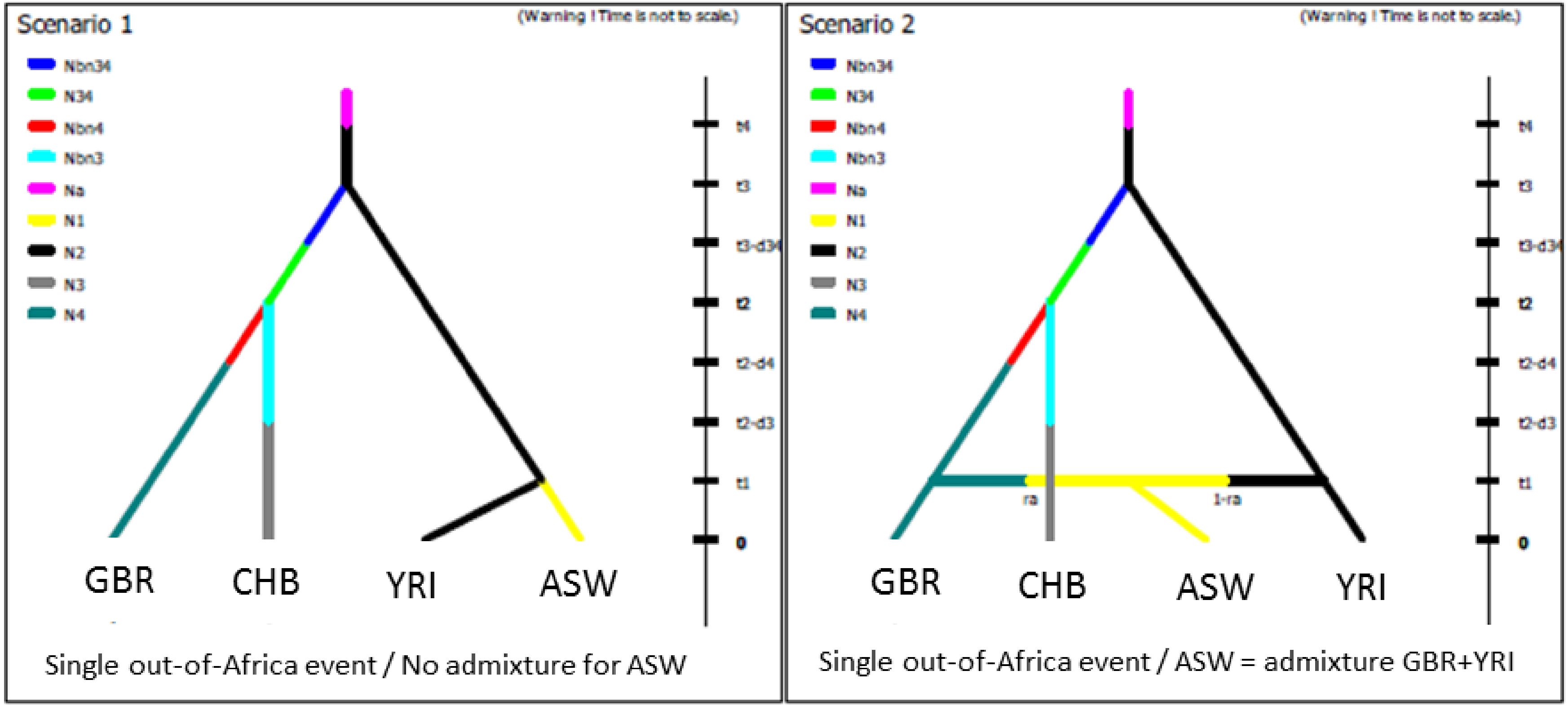
Human populations example

50,000 SNP markers genotyped in four Human populations: Yoruba (Africa), Han (East Asia), British (Europe) and American individuals of African Ancestry; 30 individuals per population.

We compared six scenarios of evolution which differ from each other by one ancient and one recent historical events:

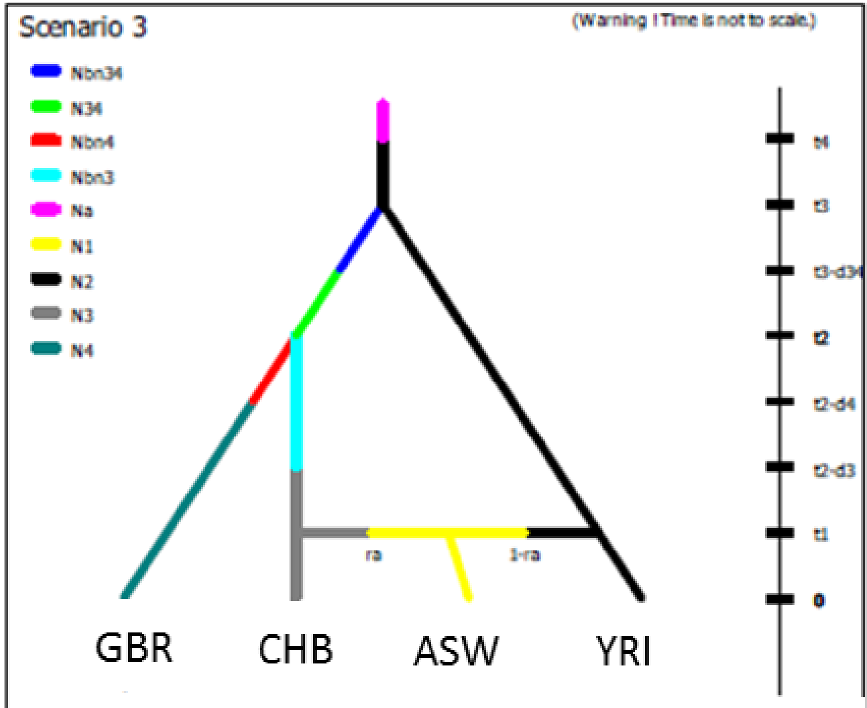
A) a single out-of-Africa colonization event giving an ancestral out-of-Africa versus two independent out-of-Africa colonization events;

B) the possibility of a recent genetic admixture of Americans of African origin with their African ancestors and individuals of European or East Asia origins.

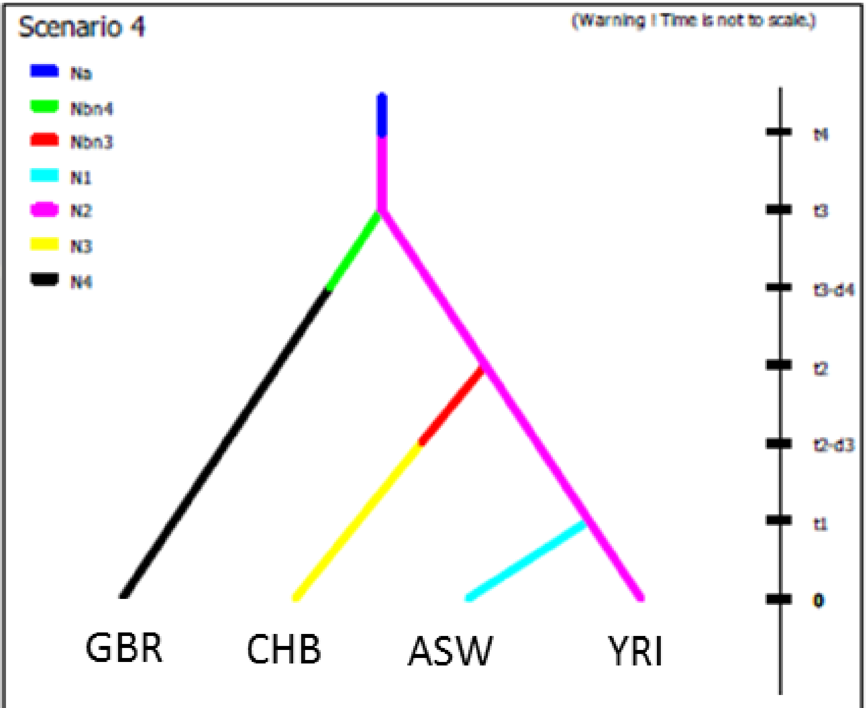


Single out-of-Africa event / No admixture for ASW

Single out-of-Africa event / ASW = admixture GBR+YRI

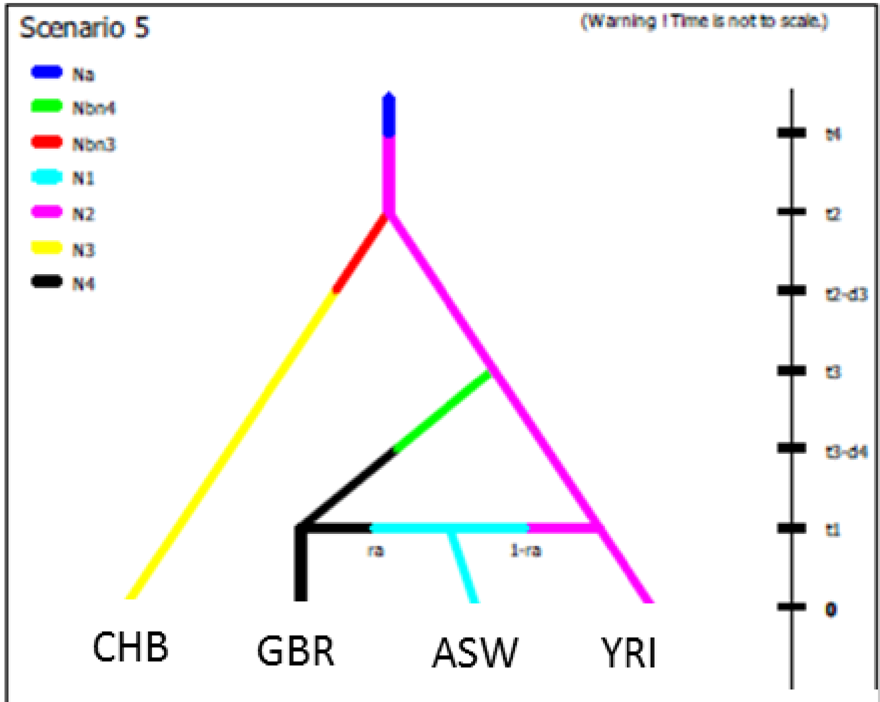


Single out-of-Africa events / ASW = admixture CHB+YRI



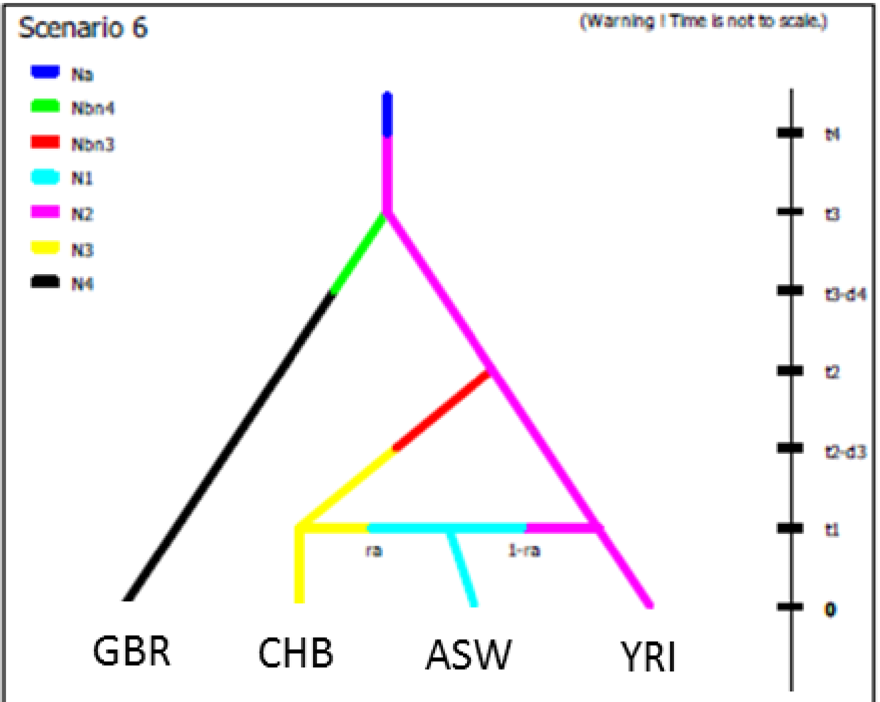
Double out-of-Africa event / No admixture for ASW

Scenario 5



Double out-of-Africa event / ASW = admixture GBR+YRI

Scenario 6



Double out-of-Africa event / ASW = admixture CHB+YRI

Outline

A - Population genetics models

A.1 - Data

A.2 - Sample genealogies

A.3 - Mutation process

A.4 - Inferential difficulties

B - ABC methods for model choice

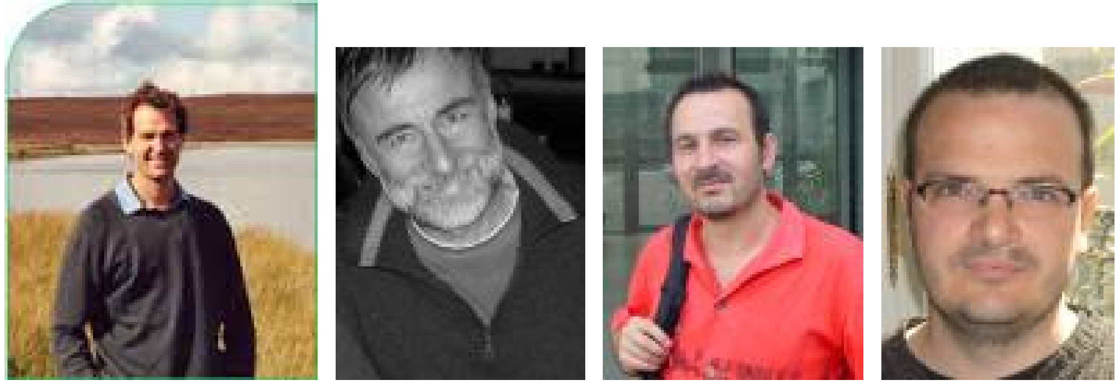
B.1 - Bayesian model choice

B.2 - Standard ABC model choice

B.3 - The use of machine learning procedures

C - Human populations example

A - Population genetics models



Mark Beaumont, Jean-Marie Cornuet, Arnaud Estoup, Mathieu Gautier

Raphaël Leblois, Pierre Pudlo, François Rousset, Mohammed Sedki

A.1 Data

The dataset consists of different samples.

Each sample corresponds to a population.

We consider D populations, $Pop1, \ldots, Pop$D, the sample size of population Popi is denoted by n_i.

We assimilate a diploid individual to two haploid individuals.

For each individual, **we consider a large number of locus on the genome**.

For these loci, the DNA sequence can vary for an individual to another due to mutations: genetic polymorphism.

Different types of loci: **microsatellite**, **SNP** (Single Nucleotide Polymorphism) or **a sequence of DNA**.

A.2 Sample genealogies

Coalescent theory (Kingman (1982), Tajima, Tavaré...)

The Kingman coalescent model describes the genealogy of a sample of genes back to the Most Recent Common Ancestor (MRCA) of the sample

The genealogy of a sample is represented by a dendrogram. Ancestral lineages are generated until the MRCA.

A coalescent event occurs when the lineages of two individuals merge at a node of the dendrogram.

Let $T_k, \dots, T_2$ be the durations between successive coalescent events.

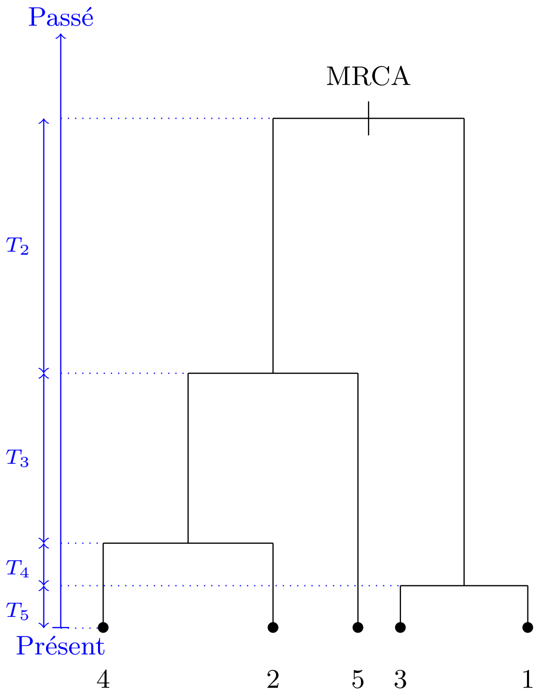
The genealogy probability distribution of k individuals is characterized by the choice of the lineages at each coalescent event and the distribution of durations between coalescent events $T_k, \dots, T_2$.

Let Ne be the effective population size.

For the Kingman coalescent, the durations between coalescent events $T_k, \dots, T_2$ are independent and T_k is distributed according to an exponential distribution with parameter $k\big(k-1\big)/(2Ne)$.

```
while $k \geq 2$ do
    1)  Generate a duration $T_k$ from an exponential distribution with parameter $\frac{k\left(k-1\right)}{2Ne}$
    2)  Add $T_k$ to the lengths of the $k$ lineages
    3)  Choose at random (uniformly) two lineages and merge them to create a node of the dendrogram
    4)  $k \leftarrow k-1$
end while
```

Five individuals from a closed population at equilibrium



Several structured populations

We consider an evolutionary scenario described by inter-populational events.

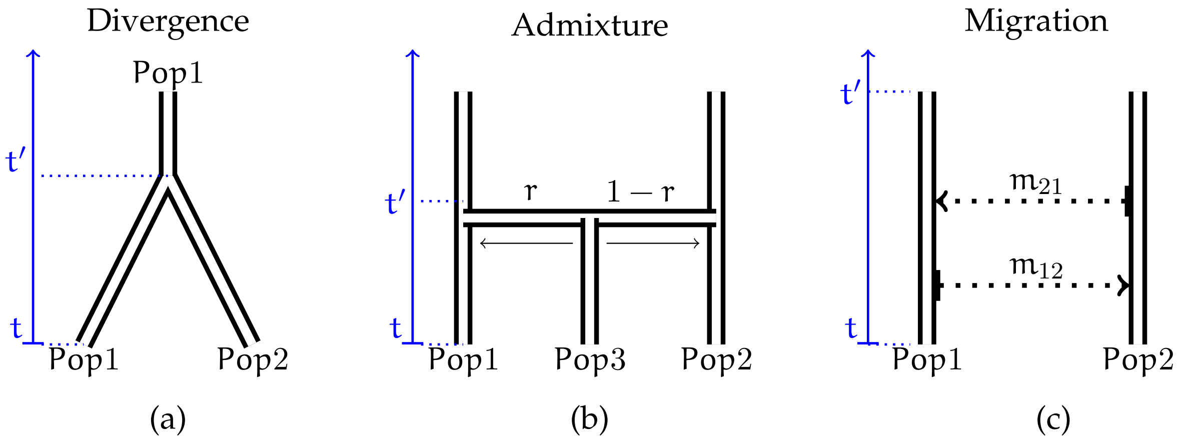
We combine these events with the Kingman coalescent which describes the intra-populational genealogies.

Three inter-populational events:

- **Divergence** the fusion of two populations back in time.
- **Admixture** the split of a population in two parts.
- **Migration** move of lineages from one population to another over a fixed period.

At a divergence event, back in time, the lineages of the two populations are merged to constitute a new population.

At an admixture event, back in time, the admixture sample $Pop3$ is split at random in two parts: a lineage of $Pop3$ is associated to $Pop1$ with probability r and to $Pop2$ with probability $1 - r$ where r is the admixture rate.



a coalescent process within each pipeline

A.3 Mutation process

Position of mutations on the tree

The mutation rate per diploid individual is denoted by μ.

Conditional on the genealogy, the mutations are distributed according to a Poisson point process with intensity $\mu/2$.

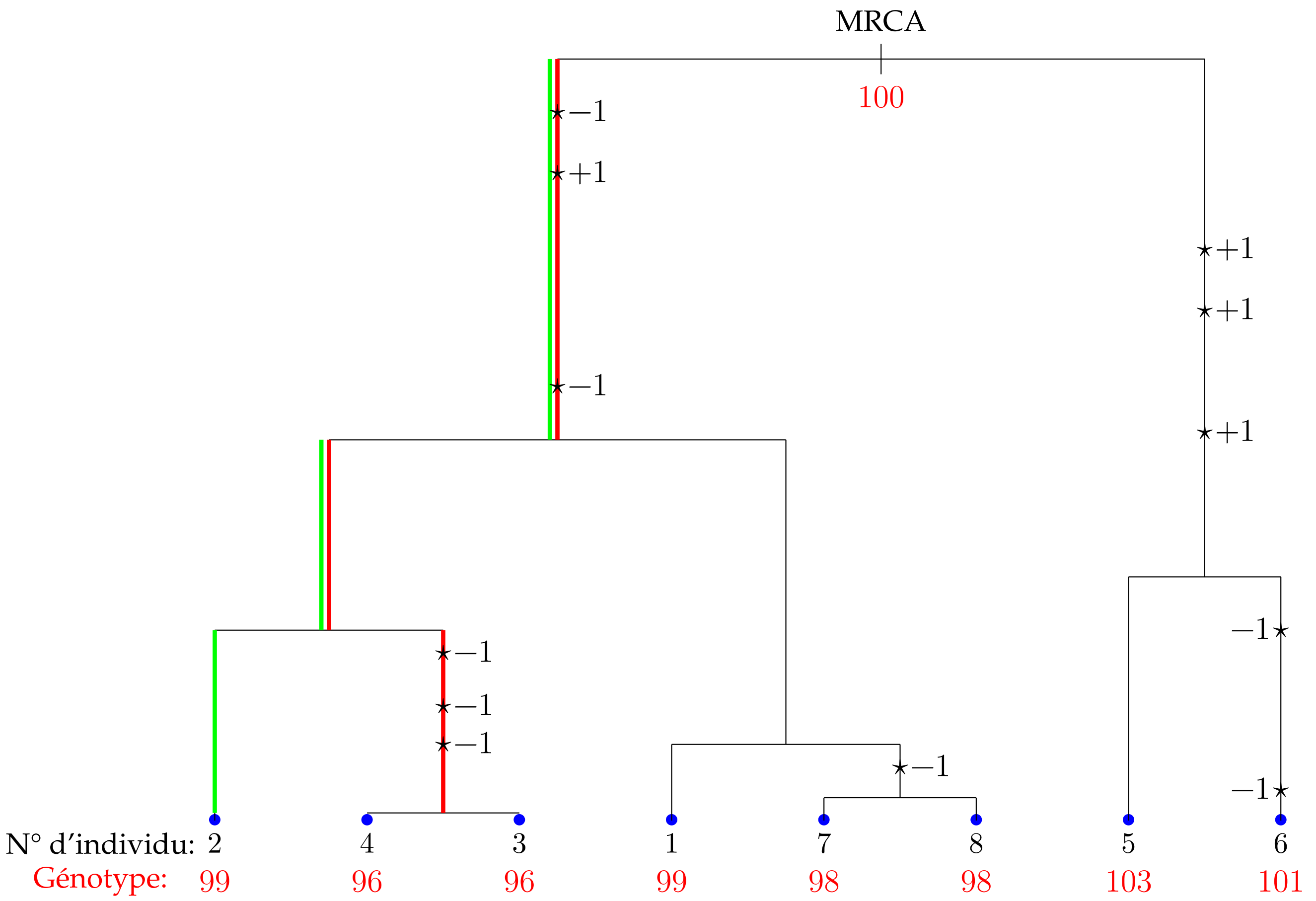
On a branch of length t, the number of mutations N is distributed according to a Poisson distribution with parameter $\mu t/2$ and the N mutations are uniformly distributed on the branch.

For microsatellite data, two mutation models: SMM (Stepwise Mutation Model) and GSM (Generalized Stepwize Mutation Model), symmetric random walks.

SNP loci have low mutation rates, we consider that polymorphism at such loci results from a single mutation.

For SNP data, a single mutation event is put at random on one branch of the genealogy, the branch being chosen with a probability proportional to its length.

To generate the genotypes of a sample at a given locus, we just have to modify the genotype of the MRCA along the genealogy.


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 Inferential difficulties

Each model is characterized by a set of parameters $\boldsymbol{\theta}$ historical (divergence times, admixture times, ...), demographic (effective population sizes, admixture rates, migration rates, ...) and genetic (mutation rate, ...).

The goal is to estimate these parameters from a polymorphism dataset $\mathbf{x}$ observed at the present time.

Difficulty: we cannot calculate the likelihood of x $f(\mathbf{x}|\boldsymbol{\theta})$.

Let $f_{\boldsymbol{\theta}}(\mathcal{G})$ denote the probability density of the genealogy of genes $\mathrm{d}\mathcal{G}$.

Let $f_{\boldsymbol{\theta}}(\mathcal{M}|\mathcal{G})$ denote the probability density of the mutation process $\mathcal{M}$ given the genealogy $\mathcal{G}$.

The likelihood is given by

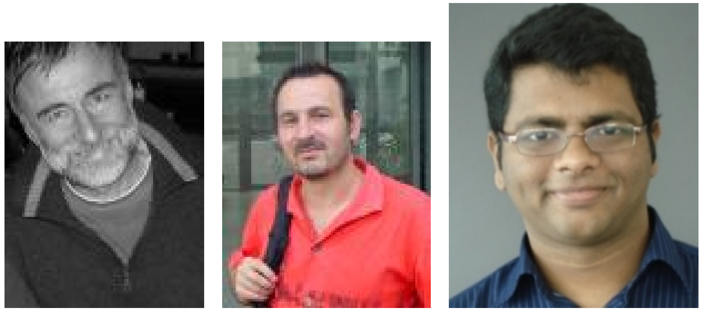
$$\ell\big(\mathbf{x}|\phi\big) = \prod_{i \in \{locus\}} \int_{\mathcal{M}_i \to \mathbf{x}_i} f_{\boldsymbol{\theta}}\big(\mathcal{M}_i\big|\mathcal{G}_i\big) f_{\boldsymbol{\theta}}\big(\mathcal{G}_i\big)\, \mathrm{d}\mathcal{G}_i\, \mathrm{d}\mathcal{M}_i, \tag{1}$$

where $\mathbf{x}_i$ is the data at locus i and $\mathcal{M}_i \to \mathbf{x}_i$ is the set of genotypes on the dendrogram compatible with $\mathbf{x}_i$.

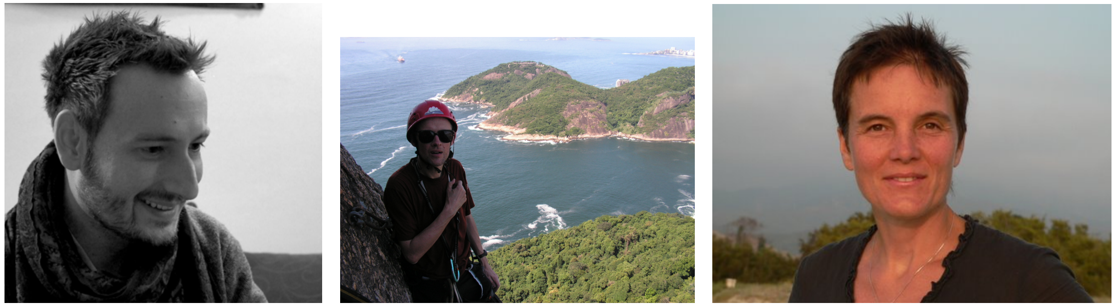
That is a very high-dimensional integral with discrete (such as the genotype) and continuous (such as the length of branches) parts.

Despite the simplicity of the Kingman coalescent and the mutation processes, we cannot expect any simplification in the calculation of the likelihood.

B - Approximate Bayesian Computation methods for model choice



Jean-Marie Cornuet, Arnuad Estoup, Natesh Pillai



Pierre Pudlo, Christian Robert, Judith Rousseau

B.1 - Bayesian model choice

J models in competition

A model is characterized by a likelihood function $f_k(\mathbf{y}|\boldsymbol{\theta}_k)$ and a prior distribution on the parameter $\boldsymbol{\theta}_k \in \Theta_k$.

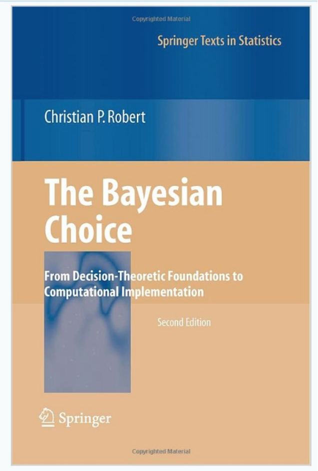
Prior probabilities in the model space are defined.

The posterior distribution in the model space is such that

$$\mathbb{P}^{\pi}\left(\mathcal{M} = k|\mathbf{y}\right) \propto \mathbb{P}(\mathcal{M} = k) \int_{\Theta_k} f_k(\mathbf{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)\, \mathbf{d}\boldsymbol{\theta}_k \,.$$

Some computational difficulties:

- How to approximate the evidences?
- When the number of models in consideration is huge, how to explore the models's space?
- How to proceed when the calculation of the likelihood in intractable?



Copyrighted Material

Springer Texts in Statistics

Christian P. Robert

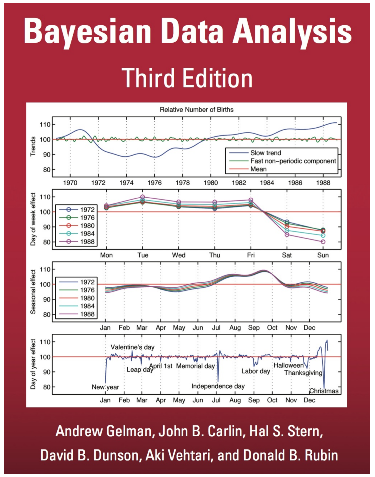
The Bayesian Choice

From Decision-Theoretic Foundations to Computational Implementation

Second Edition

Springer

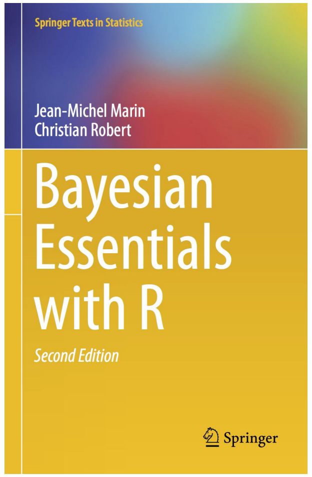
Copyrighted Material



Bayesian Data Analysis

Third Edition

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

B.2 - Standard ABC model choice procedure

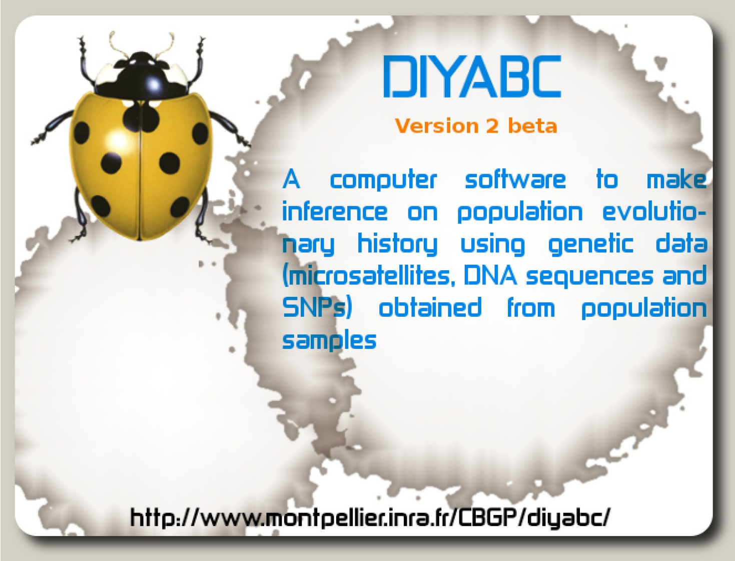
Let $N = \lfloor \alpha M \rfloor$.

```
1) For $i = 1, \ldots, M$:
    a) Generate $m_i$ from the prior $\pi(\mathcal{M} = m)$,
    b) Generate $\boldsymbol{\theta}'_{m_i}$ from the prior $\pi_{m_i}(\cdot)$,
    c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\boldsymbol{\theta}'_{m_i})$,
    d) Calculate $d_i = \rho(\eta(\mathbf{z}), \eta(\mathbf{y}))$,
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$,
3) Return the $m_i$'s that correspond to the $N$-smallest distances.
```

Note that **Ratmann et al. (2009) PNAS** proposed methodology for testing the fit of a model without reference to other models.

Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation **Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, Estoup (2008) Bioinformatics**

DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data **Cornuet, Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin, Estoup (2014) Bioinformatics**



DIYABC

Version 2 beta

A computer software to make inference on population evolutionary history using genetic data (microsatellites, DNA sequences and SNPs) obtained from population samples

http://www.montpellier.inra.fr/CBGP/diyabc/

If $\eta(\mathbf{y})$ is a sufficient statistics for the model choice problem, 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

If not...

Lack of confidence in approximate Bayesian computation model choice

Robert, Cornuet, Marin, Pillai (2011) PNAS

Relevant statistics for Bayesian model choice

Marin, Pillai, Robert, Rousseau (2014) JRSS B

B.4 - The use of machine learning procedures

The standard ABC model choice technique corresponds to a knn approximation of the posterior probabilities!

New insights into Approximate Bayesian Computation

Biau, Cérou, Guyader (2015) Annales de l'IHP

We investigate some ABC model choice techniques that use others machine learning procedures.

Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics

Estoup, Lombaert, Marin, Guillemaud, Pudlo, Robert, Cornuet (2012) Molecular Ecology

Key points

- ABC model choice seen as learning about which model is most appropriate from a huge (reference) table
- exploiting a large number of summary statistics is not an issue for some machine learning methods intended to estimate efficient combinations

Random Forests

Technique that stemmed from Leo Breiman's bagging (or bootstrap aggregating) machine learning algorithm for both classification and regression

Bagging Predictors

Breiman (1996) Machine Learning

Improved classification performances by averaging over classification schemes of randomly generated training sets, creating a forest of CART decision trees

Random forests

Breiman (2001) Machine Learning

Breiman's solution for inducing random features in the trees of the forest:

- boostrap resampling of the dataset and
- random subseting of the covariates driving the classification at every node of each tree

Input ABC reference table involving model index and summary statistics for the associated simulated pseudo-data [possibly large collection of summary statistics (from scientific theory input to available statistical softwares, to machine-learning alternatives)]

Output a random forest classifier to infer model indexes $\widehat{m(\eta(\mathbf{y}))}$

Some theoretical guarantees for sparse problems:

Analysis of a random forest model

Biau (2012) JMLR

Consistency of random forests

Scornet, Biau, Vert (2015) The Annals of Statistics

Random forest predicts a MAP model index, from the observed dataset: the predictor provided by the forest is good enough to select the most likely model but not to derive directly the associated posterior probability

frequency of trees associated with majority model is no proper substitute to the true posterior probability

Estimate of the posterior probability of the selected model

$$\mathbb{P}[\mathcal{M} = \widehat{m(\eta(\mathbf{y}))}|\eta(\mathbf{y})]$$

random comes from $\mathcal{M}$ (bayesian)!

$$\mathbb{P}[\mathcal{M} = \widehat{m(\eta(\mathbf{y}))}|\eta(\mathbf{y})] = 1 - \mathbb{E}\left[\mathbb{I}(\mathcal{M} \neq \widehat{m(\eta(\mathbf{y}))})|\eta(\mathbf{y})\right]$$

A second random forest in regression

1) compute the value of $\mathbb{I}(\mathcal{M} \neq \widehat{m(\eta(\mathbf{z}))}$ for the trained random forest $\hat{m}$ and for all terms in the ABC reference table using the out-of-bag classifiers;

2) train a RF regression and get $\varrho(\eta(\mathbf{z})) = \widehat{\mathbb{E}\left[\mathbb{I}(\mathcal{M} \neq \widehat{m(\eta(\mathbf{z}))})|\eta(\mathbf{z})]\right]}$;

3) return $\widehat{\mathbb{P}[\mathcal{M} = \widehat{m(\eta(\mathbf{y}))}|\eta(\mathbf{y})]} = 1 - \varrho(\eta(\mathbf{y}))$.

on same reference table out-of-bag magic trick avoid overfitting!

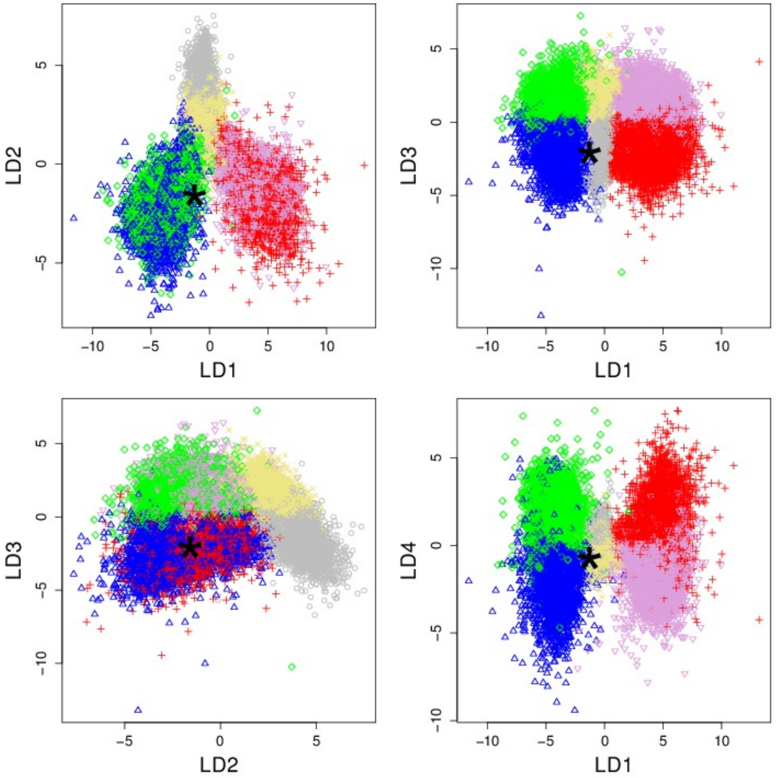
Reliable ABC model choice via random forests

Pudlo, Marin, Estoup, Cornuet, Gauthier, Robert (2015) Bioinformatics

The proposed methodology is implemented in the R package abcrf available on the CRAN.

C - Human populations example

We used all the summary statistics provided by DIYABC for SNP markers, namely 112 statistics in this setting complemented by the five LDA axes as additional statistics.



Estimated prior error rates for classification methods and three sizes of reference table

Classification method	**Prior error rates** (%)		
trained on	$N_{\text{ref}} = 10,000$	$N_{\text{ref}} = 20,000$	$N_{\text{ref}} = 50,000$
LDA	9.91	9.97	10.03
standard ABC	23.18	20.55	17.76
standard ABC with LDA axes	6.29	5.76	5.70
local logistic reg.	6.85	6.42	6.07
RF	8.84	7.32	6.34
RF with LDA axes	5.01	4.66	4.18

ABC-RF on the Human dataset selects Scenario 2 as the forecasted scenario.

Considering previous population genetics studies in the field, it is not surprising that this scenario was selected.

It includes a single out-of-Africa colonization event giving an ancestral out-of-Africa population with a secondarily split into one European and one East Asian population lineage and a recent genetic admixture of Americans of African origin with their African ancestors and European individuals.

This selection is associated with a high confidence level as we got an estimate of the posterior probability of scenario 2 equal to 0.998.