Computing Bayesian posterior with empirical likelihood in population genetics

Pierre Pudlo

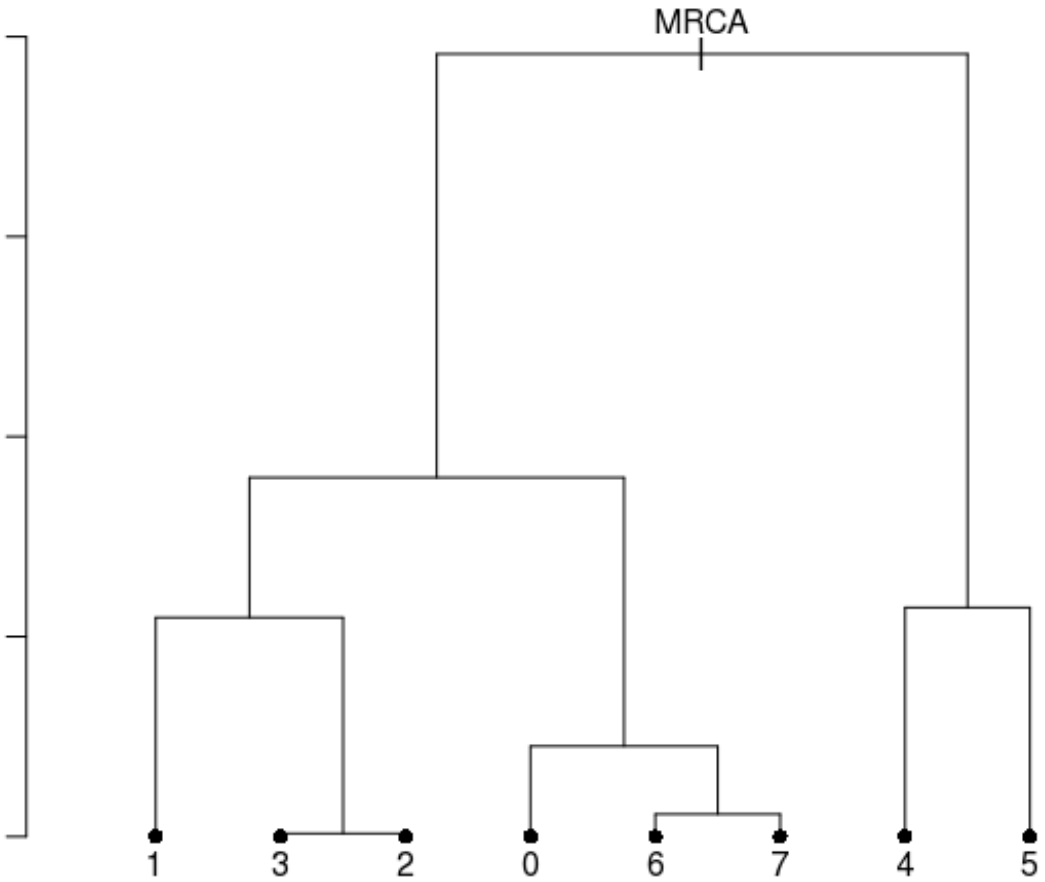
INRA & U. Montpellier 2

MCEB, June 2012

Table of contents

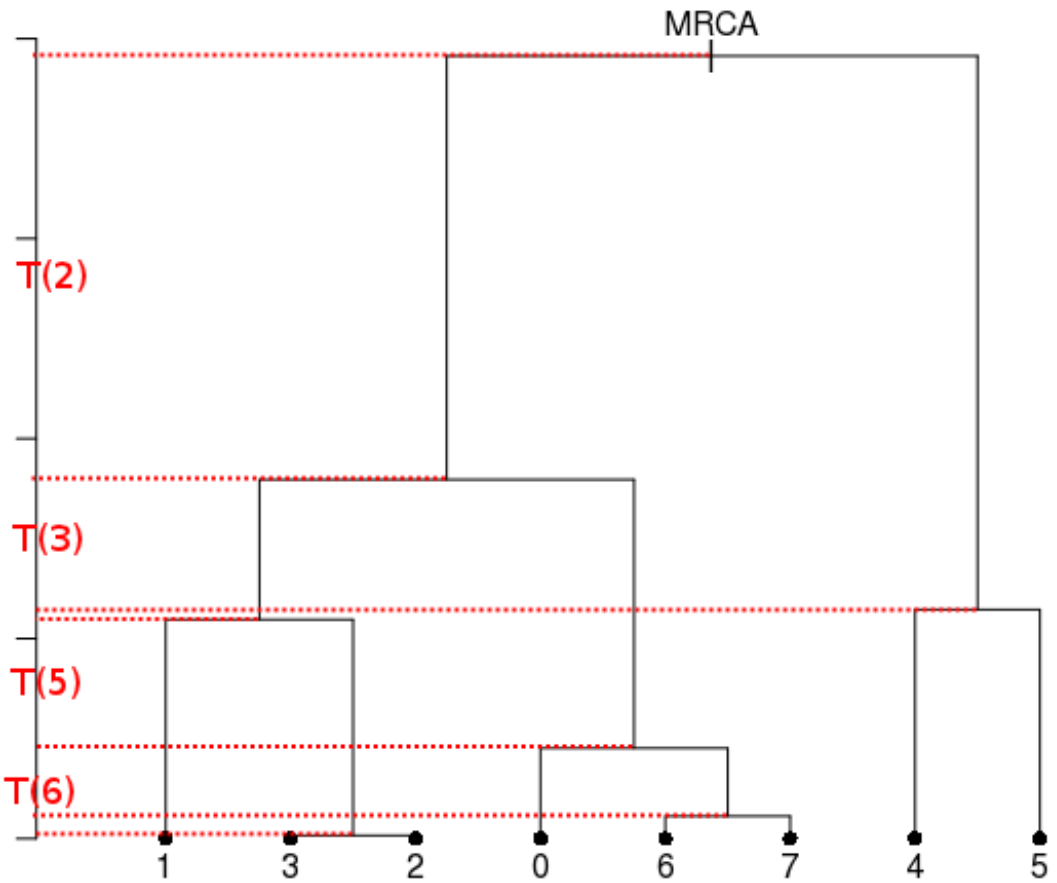
Table of contents

Neutral model at a given microsatellite locus, in a closed panmictic population at equilibrium



Sample of 8 genes

Neutral model at a given microsatellite locus, in a closed panmictic population at equilibrium

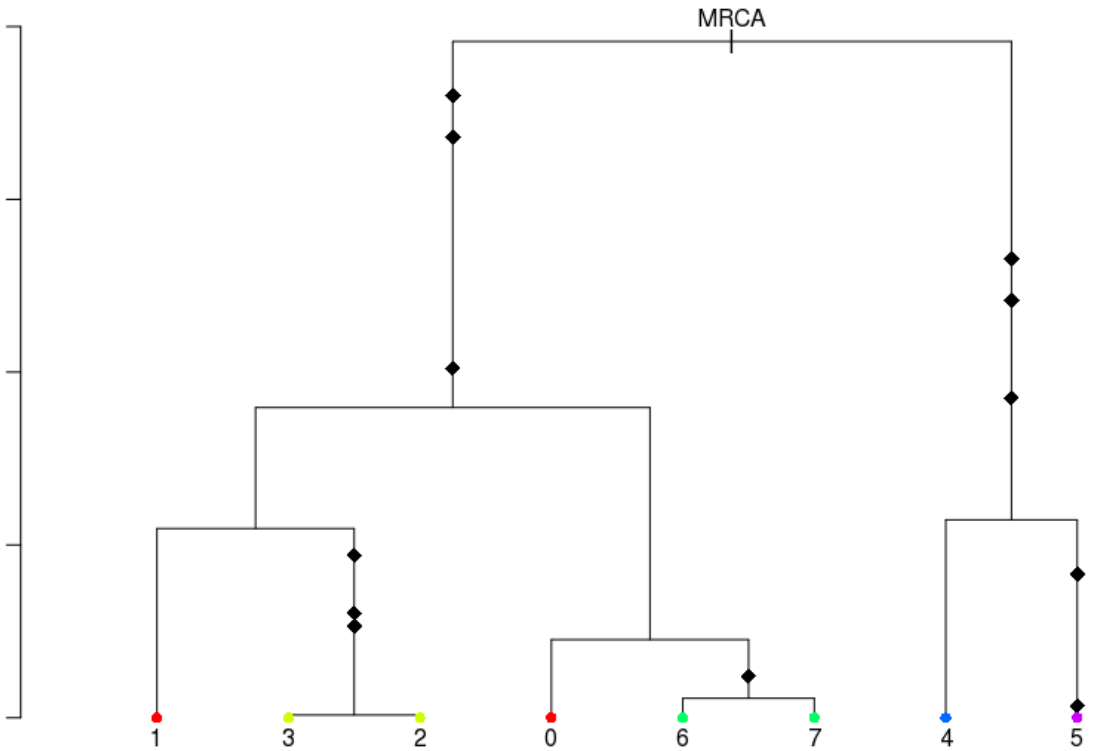


Kingman's genealogy

When time axis is normalized,

$T(k) \sim \text{Exp}(k(k-1)/2)$

Neutral model at a given microsatellite locus, in a closed panmictic population at equilibrium



Kingman's genealogy

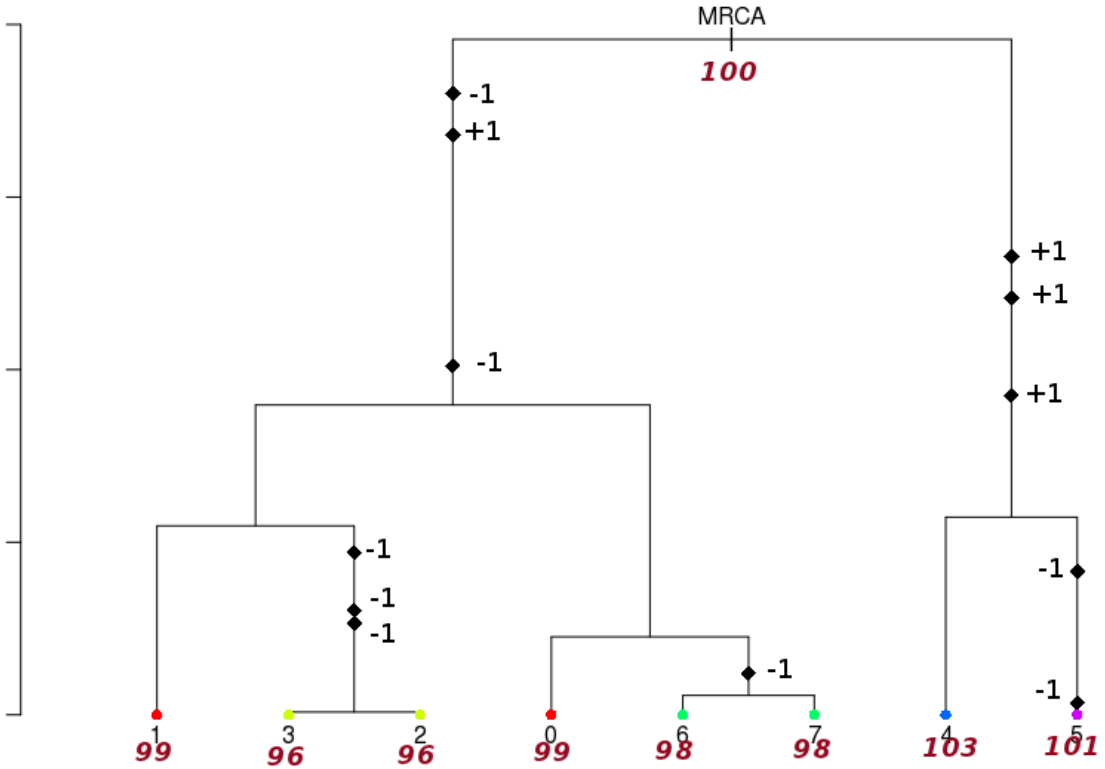
When time axis is normalized,

$T(k) \sim \text{Exp}(k(k-1)/2)$

Mutations according to the Simple stepwise Mutation Model (SMM)

- date of the mutations $\sim$ Poisson process with intensity $\theta/2$ over the branches

Neutral model at a given microsatellite locus, in a closed panmictic population at equilibrium



Observations: leafs of the tree

$$\hat{\theta} = ?$$

Kingman's genealogy

When time axis is normalized,

$T(k) \sim \text{Exp}(k(k-1)/2)$

Mutations according to the Simple stepwise Mutation Model (SMM)

- date of the mutations $\sim$ Poisson process with intensity $\theta/2$ over the branches
- MRCA $= 100$
- independent mutations: ± 1 with pr. $1/2$

Much more interesting models...

- **several independent loci**
  Independent gene genealogies and mutations
- **different populations**
  linked by an evolutionary scenario made of divergences, admixtures, migrations between populations, etc.
- **larger sample size**
  usually between 50 and 100 genes

A typical evolutionary scenario:

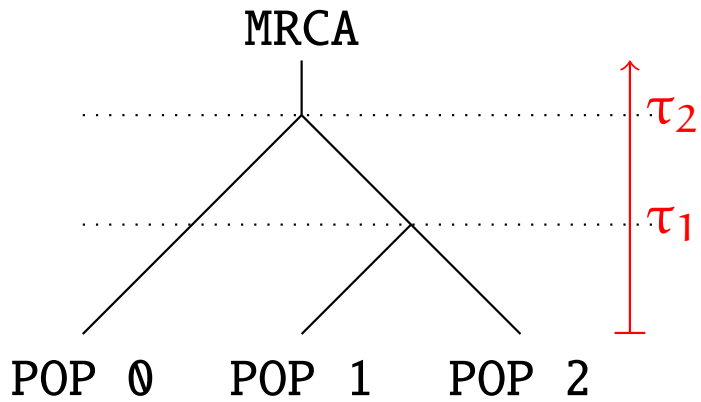


Table of contents

When the likelihood is not completely known

- **Hidden Markov and other dynamic models:** latent process which is not observed

$\hookrightarrow$ Classical answer: Markov chain Monte Carlo,...

- **Population genetics:** the whole gene genealogy is unobserved Likelihood is an integral over
  - all possible gene genealogies
  - all possible mutations along the genealogies

$\hookrightarrow$ Classical answer: Approximate Bayesian computation (ABC)

ABC in a nutshell

Posterior distribution is the conditional distribution of

$$\pi(\phi)\ell(x|\phi) \quad (*)$$

knowing that $x = x_{obs}$

Methodology

Draw a (large) set of particles (ϕ_i, x_i) from $(*)$ and use a nonparametric estimate of the conditional density

$$\pi(\phi|x_{obs}) \propto \pi(\phi)\ell(x_{obs}|\phi)$$

Seminal papers

- Tavaré, Balding, Griffith and Donnelly (1997, Genetics)
- Pritchard, Seielstad, Perez-Lezuan, Feldman (1999, Molecular Biology and Evolution)

ABC in a nutshell

Posterior distribution is the conditional distribution of

$$\pi(\phi)\ell(x|\phi) \quad (*)$$

knowing that $x = x_{obs}$

Methodology

Draw a (large) set of particles (ϕ_i, x_i) from $(*)$ and use a nonparametric estimate of the conditional density

$$\pi(\phi|x_{obs}) \propto \pi(\phi)\ell(x_{obs}|\phi)$$

Shortcomings.

- *time consuming* – If simulation of the latent process is not straightforward

ABC in a nutshell

Posterior distribution is the conditional distribution of

$$\pi(\phi)\ell(x|\phi) \quad (*)$$

knowing that $x = x_{obs}$

Methodology

Draw a (large) set of particles (ϕ_i, x_i) from $(*)$ and use a nonparametric estimate of the conditional density

$$\pi(\phi|x_{obs}) \propto \pi(\phi)\ell(x_{obs}|\phi)$$

Shortcomings.

- *time consuming* – If simulation of the latent process is not straightforward
- *curse of dimensionality vs. loss of information* –
  - If x lies in a high dimensional space $\mathscr{X}$ (often), we are unable to estimate of the conditional density

ABC in a nutshell

Posterior distribution is the conditional distribution of

$$\pi(\phi)\ell(x|\phi) \quad (*)$$

knowing that $x = x_{obs}$

Methodology

Draw a (large) set of particles (ϕ_i, x_i) from $(*)$ and use a nonparametric estimate of the conditional density

$$\pi(\phi|\eta(x_{obs})) \propto \pi(\phi) \int_{x\,:\,\eta(x)=\eta(x_{obs})} \ell(x|\phi)\, dx$$

Shortcomings.

- *time consuming* – If simulation of the latent process is not straightforward
- *curse of dimensionality vs. loss of information* –
  - If x lies in a high dimensional space $\mathscr{X}$ (often), we are unable to estimate of the conditional density
  - Hence, we project the (observed and simulated) datasets on a space with smaller dimension (trough summary statistics) $\eta : \mathscr{X} \to \mathbb{R}^d$ (summary statistics)

Curse of dimensionality

Assume that

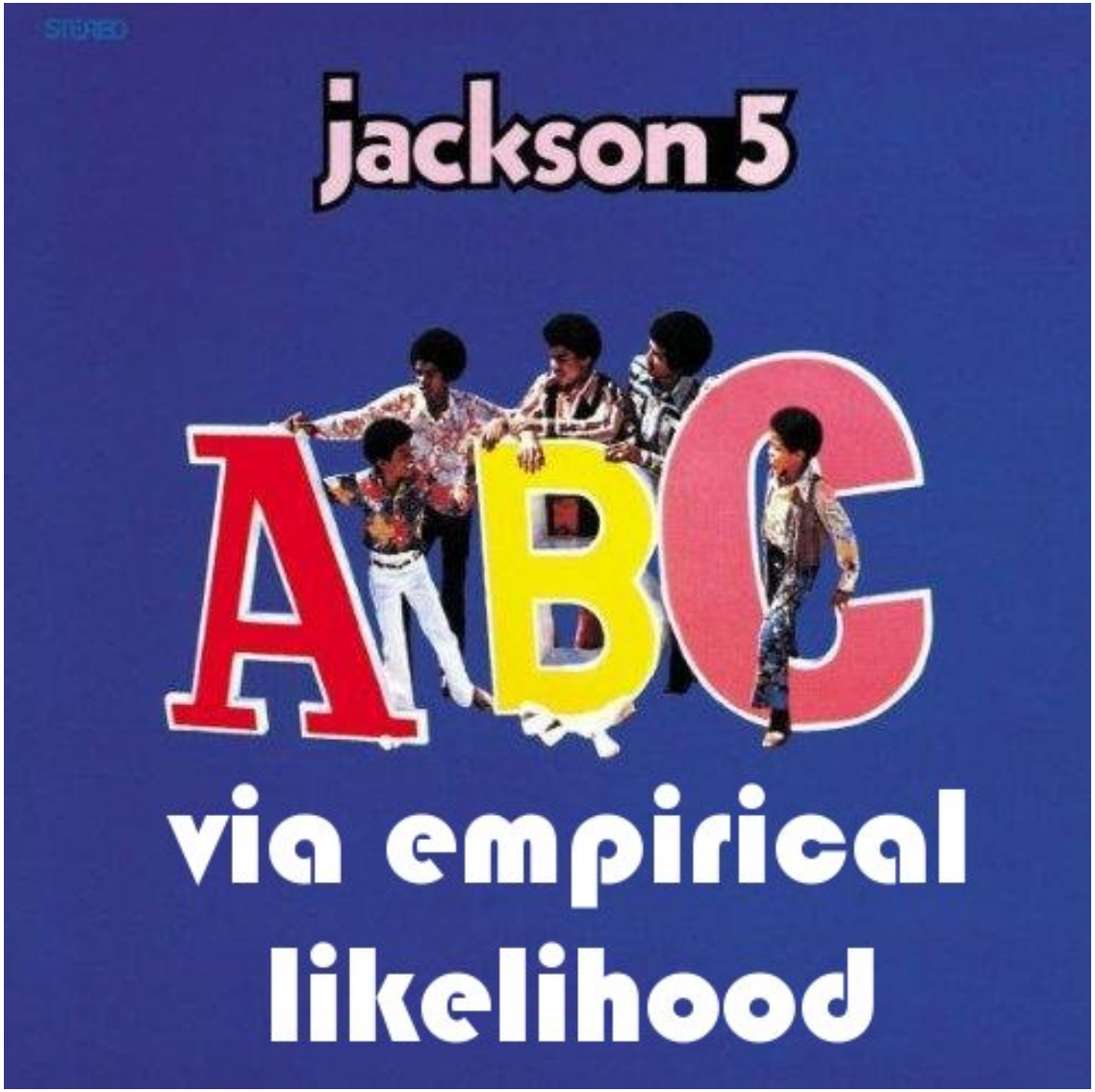
- the simulated summary statistics $\eta(x_1), \ldots, \eta(x_N)$
- the observed summary statistics $\eta(x_{obs})$

are iid, with uniform law on $[0,1]^d$

Let $d_\infty(d,N) = \mathbb{E}\left[\min_{i=1,\ldots,N} \left\|\eta(x_{obs}) - \eta(x_i)\right\|_\infty\right]$

	$N = 100$	$N = 1,000$	$N = 10,000$	$N = 100,000$
$\delta_\infty(1,N)$	0.0025	0.00025	0.000025	0.0000025
$\delta_\infty(2,N)$	$\geqslant 0.033$	$\geqslant 0.01$	$\geqslant 0.0033$	$\geqslant 0.001$
$\delta_\infty(10,N)$	$\geqslant 0.28$	$\geqslant 0.22$	$\geqslant 0.18$	$\geqslant 0.14$
$\delta_\infty(200,N)$	$\geqslant 0.48$	$\geqslant 0.48$	$\geqslant 0.47$	$\geqslant 0.46$

Table of contents


jackson 5
ABC
via empirical
likelihood

Empirical likelihood (EL)

Owen (1988, Biometrika), Owen (2001, Chapman & Hall)

Assume that the dataset x is composed of n independent replicates $x = (x_1, \ldots, x_n)$ of some $X \sim F$

Generalized moment condition model

The law F of X satisfy

$$\mathbb{E}_F\left[h(X, \phi)\right] = 0,$$

where h is a known function, and ϕ an unknown parameter

Empirical likelihood

$$L_{el}(\phi|x) = \max_p \prod_{i=1}^{n} p_i$$

for all p such that $0 \leqslant p_i \leqslant 1$, $\sum p_i = 1$, $\sum_i p_i h(x_i, \phi) = 0$.

Raw ABC_{el}sampler

We act as if EL was an exact likelihood

---

for $i = 1 \to N$ **do**
  generate ϕ_i from the prior distribution $\pi(\cdot)$
  set the weight $\omega_i = L_{el}(\phi_i | x_{obs})$
end for
return (ϕ_i, ω_i), $i = 1, \ldots, N$

- The output is sample of parameters of size N with associated weights
- Performance of the output evaluated through effective sample size

$$ESS = 1 \Big/ \sum_{i=1}^{N} \left\{ \omega_i \Big/ \sum_{j=1}^{N} \omega_j \right\}^2$$

- Other classical sampling algorithms might be adapted to use EL. We resorted the adaptive multiple importance sampling (AMIS) of Cornuet *et al.* (Scandinavian J. of Statis.) to speed up computations

Moment condition in population genetics?

EL does not require a fully defined and often complex (hence debatable) parametric model.

Main difficulty

Derive a constraint

$$\mathbb{E}_F\left[h(X,\phi)\right] = 0,$$

on the parameters of interest ϕ when X is the allelic states of our sample of individuals at a given locus

E.g., in phylogeography, ϕ is composed of

- dates of splits of populations,
- ratio of population sizes,
- mutation rates, etc.

None of them are moments of the distribution of the allelic states of the sample

Moment condition in population genetics?

EL does not require a fully defined and often complex (hence debatable) parametric model.

Main difficulty

Derive a constraint

$$\mathbb{E}_F\left[h(X,\phi)\right] = 0,$$

on the parameters of interest ϕ when X is the allelic states of our sample of individuals at a given locus

E.g., in phylogeography, ϕ is composed of

- dates of splits of populations,
- ratio of population sizes,
- mutation rates, etc.

None of them are moments of the distribution of the allelic states of the sample

$\hookrightarrow$ $h =$ pairwise composite scores whose zero is the pairwise maximum likelihood estimator

Pairwise composite likelihood?

The intra-locus pairwise likelihood

$$\ell_2(\mathbf{x}_k|\phi) = \prod_{i<j} \ell_2(x_k^i, x_k^j|\phi)$$

with $x_k^1, \ldots, x_k^n$: allelic states of the gene sample at the k-th locus

The pairwise score function

$$\nabla_\phi \log \ell_2(\mathbf{x}_k|\phi) = \sum_{i<j} \nabla_\phi \log \ell_2(x_k^i, x_k^j|\phi)$$

Composite likelihoods are often much more narrow than the distribution of the model

Safe with EL because we only use position of its mode

Pairwise likelihood: a simple case

Assumptions

- sample $\subset$ closed, panmictic population at equilibrium
- marker: microsatellite
- mutation rate: $\theta/2$

if x_k^i et x_k^j are two genes of the sample,

$\ell_2(x_k^i, x_k^j|\theta)$ depends only on $\delta = x_k^i - x_k^j$

Pairwise likelihood: a simple case

Assumptions

- sample $\subset$ closed, panmictic population at equilibrium
- marker: microsatellite
- mutation rate: $\theta/2$

$$\ell_2(\delta|\theta) = \frac{1}{\sqrt{1+2\theta}}\rho\,(\theta)^{|\delta|}$$

with

$$\rho(\theta) = \frac{\theta}{1+\theta+\sqrt{1+2\theta}}$$

if x_k^i et x_k^j are two genes of the sample,

$\ell_2(x_k^i, x_k^j|\theta)$ depends only on $\delta = x_k^i - x_k^j$

Pairwise likelihood: a simple case

Assumptions

- sample $\subset$ closed, panmictic population at equilibrium
- marker: microsatellite
- mutation rate: $\theta/2$

if x_k^i et x_k^j are two genes of the sample,

$\ell_2(x_k^i, x_k^j|\theta)$ depends only on $\delta = x_k^i - x_k^j$

$$\ell_2(\delta|\theta) = \frac{1}{\sqrt{1+2\theta}}\rho(\theta)^{|\delta|}$$

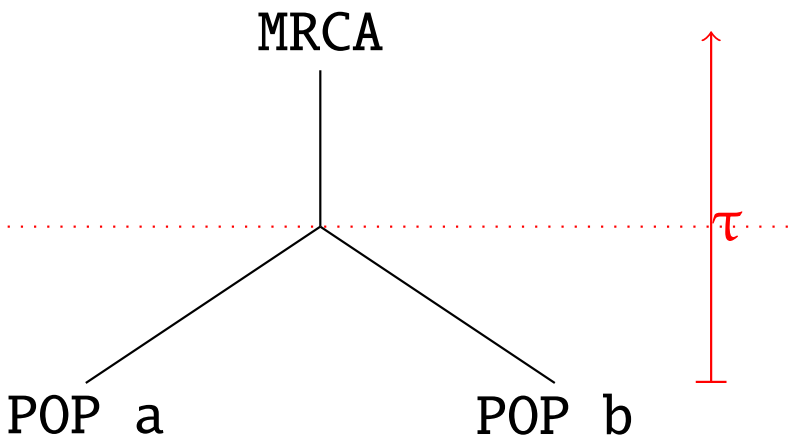
with

$$\rho(\theta) = \frac{\theta}{1+\theta+\sqrt{1+2\theta}}$$

Pairwise score function

$$\partial_\theta \log \ell_2(\delta|\theta) = -\frac{1}{1+2\theta} + \frac{|\delta|}{\theta\sqrt{1+2\theta}}$$

Pairwise likelihood: 2 diverging populations



Assumptions

- τ: divergence date of pop. a and b
- $\theta/2$: mutation rate

Let x_k^i and x_k^j be two genes coming resp. from pop. a and b
Set $\delta = x_k^i - x_k^j$.

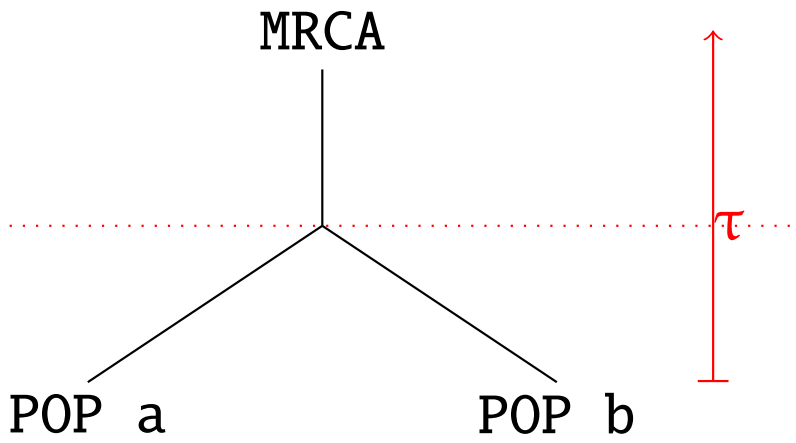
Then $\ell_2(\delta|\theta,\tau) =$

$$\frac{e^{-\tau\theta}}{\sqrt{1+2\theta}} \sum_{k=-\infty}^{+\infty} \rho(\theta)^{|k|} I_{\delta-k}(\tau\theta).$$

where

$I_n(z)$ nth-order modified Bessel function of the first kind

Pairwise likelihood: 2 diverging populations



Assumptions

- τ: divergence date of pop. a and b
- $\theta/2$: mutation rate

Let x_k^i and x_k^j be two genes coming resp. from pop. a and b
Set $\delta = x_k^i - x_k^j$.

A 2-dim score function

$$\partial_\tau \log \ell_2(\delta|\theta,\tau) =$$
$$-\theta + \frac{\theta}{2}\frac{\ell_2(\delta-1|\theta,\tau)+\ell_2(\delta+1|\theta,\tau)}{\ell_2(\delta|\theta,\tau)}$$

$$\partial_\theta \log \ell_2(\delta|\theta,\tau) =$$
$$-\tau - \frac{1}{1+2\theta} +$$
$$\frac{\tau}{2}\frac{\ell_2(\delta-1|\theta,\tau)+\ell_2(\delta+1|\theta,\tau)}{\ell_2(\delta|\theta,\tau)} +$$
$$\frac{q(\delta|\theta,\tau)}{\ell_2(\delta|\theta,\tau)}$$

where

$$q(\delta|\theta,\tau) :=$$
$$\frac{e^{-\tau\theta}}{\sqrt{1+2\theta}}\frac{\rho'(\theta)}{\rho(\theta)}\sum_{k=-\infty}^{\infty}|k|\rho(\theta)^{|k|}I_{\delta-k}(\tau\theta)$$

Recap

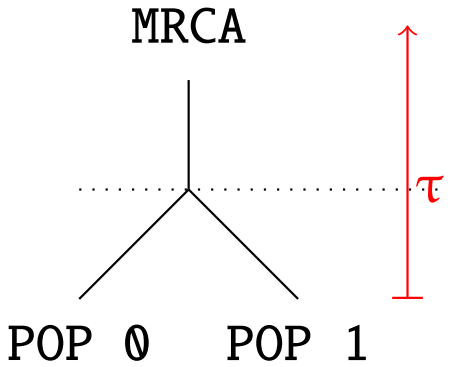
Three kinds of likelihood:

- **True likelihood**: given by the **model** (evolutionary scenario & Kingman's coalecent)
  $\hookrightarrow$ cannot compute
- **Pairwise composite likelihood**: act as if each pair of genes was independent of the other ones
  $\hookrightarrow$ its maximum provides as "good" approximation of the MLE
- **Empirical likelihood**: a way to profile the likelihood **from the data**, using generalized moment conditions
  $\hookrightarrow$ generalized moment condition in population genetics $=$ pairwise composite scores (whose zero is the pairwise composite maximum likelihood)

Table of contents

A first experiment

Evolutionary scenario:



Dataset:

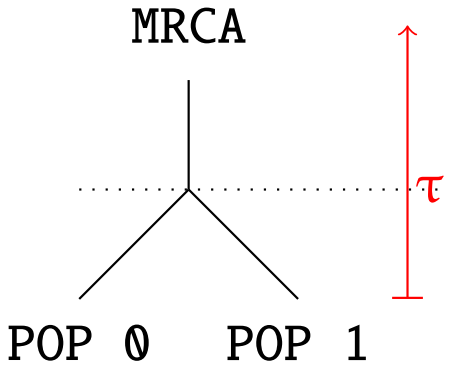
- 50 genes per populations,
- 100 microsat. loci

Assumptions:

- N_e identical over all populations
- $\phi = (\log_{10}\theta, \log_{10}\tau)$
- uniform prior over $(-1., 1.5) \times (-1., 1.)$

A first experiment

Evolutionary scenario:



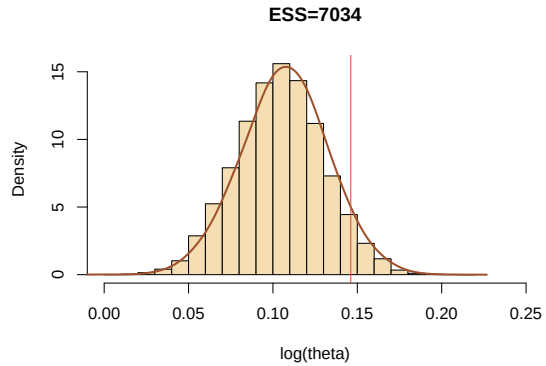
Dataset:

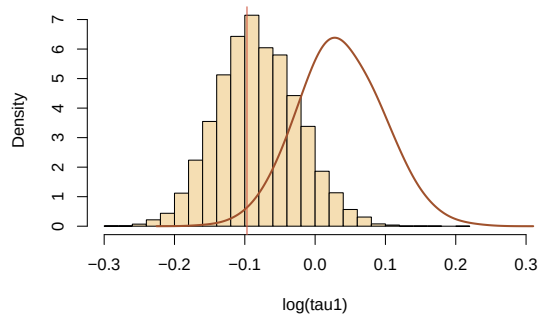
- 50 genes per populations,
- 100 microsat. loci

Assumptions:

- N_e identical over all populations
- $\phi = (\log_{10}\theta, \log_{10}\tau)$
- uniform prior over $(-1., 1.5) \times (-1., 1.)$

Comparison of the original ABC with ABC_{el}





histogram = ABC_{el}
curve = original ABC
vertical line = “true” parameter

ABC vs. ABC_{el} on 100 replicates of the 1st experiment

Accuracy:

	$\log_{10}\theta$		$\log_{10}\tau$	
	ABC	ABC_{el}	ABC	ABC_{el}
(1)	0.097	0.094	0.315	0.117
(2)	0.68	0.81	1.0	0.80

(1) Root Mean Square Error of the posterior mean

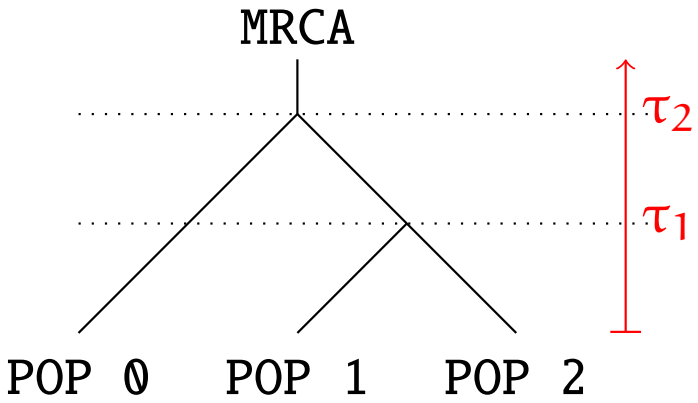
(2) Coverage of the credibility interval of probability 0.8

Computation time: on a recent 6-core computer (C++/OpenMP)

- ABC ≈ 4 hours
- $ABC_{el} \approx 2$ minutes

Second experiment

Evolutionary scenario:



Dataset:

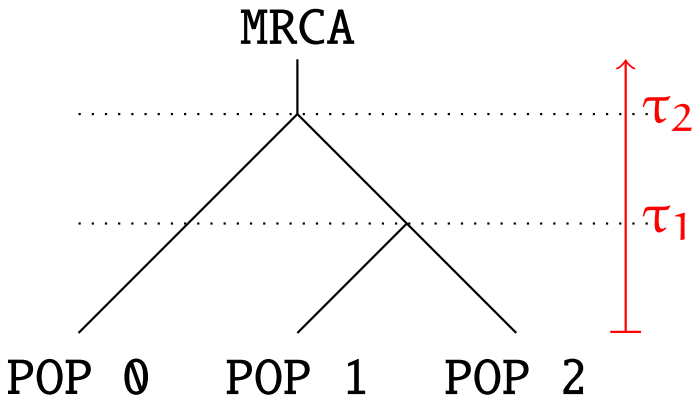
- 50 genes per populations,
- 100 microsat. loci

Assumptions:

- N_e identical over all populations
- $\phi = (\log_{10}\theta, \log_{10}\tau_1, \log_{10}\tau_2)$
- non-informative prior

Second experiment

Evolutionary scenario:



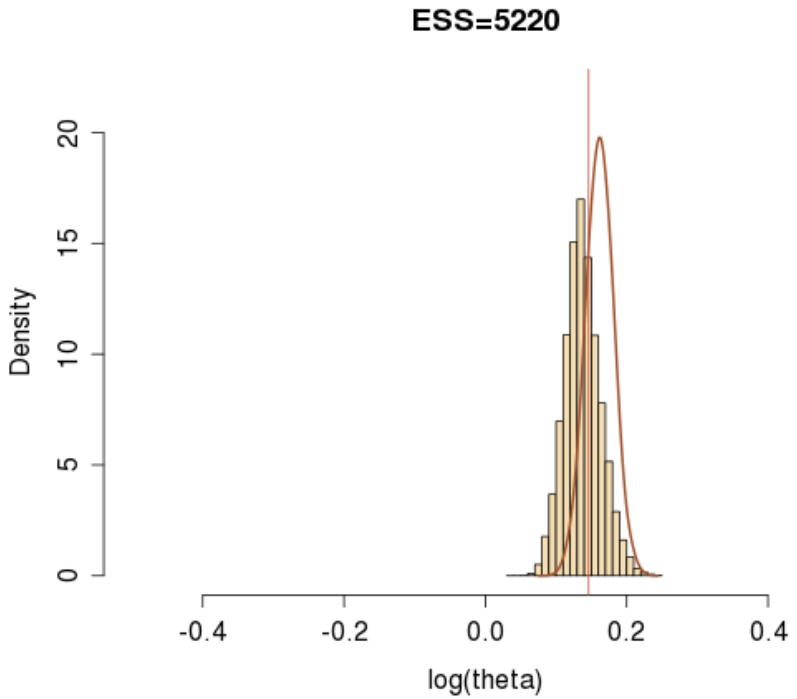
Dataset:

- 50 genes per populations,
- 100 microsat. loci

Assumptions:

- N_e identical over all populations
- $\phi = (\log_{10}\theta, \log_{10}\tau_1, \log_{10}\tau_2)$
- non-informative prior

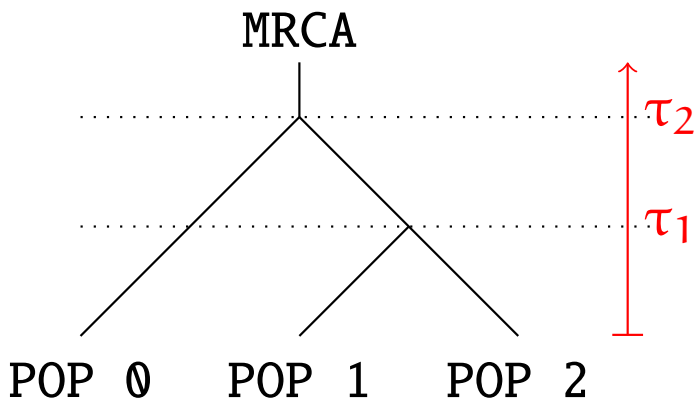
Comparison of the original ABC with ABC_{el}



histogram = ABC_{el}
curve = original ABC
vertical line = “true” parameter

Second experiment

Evolutionary scenario:



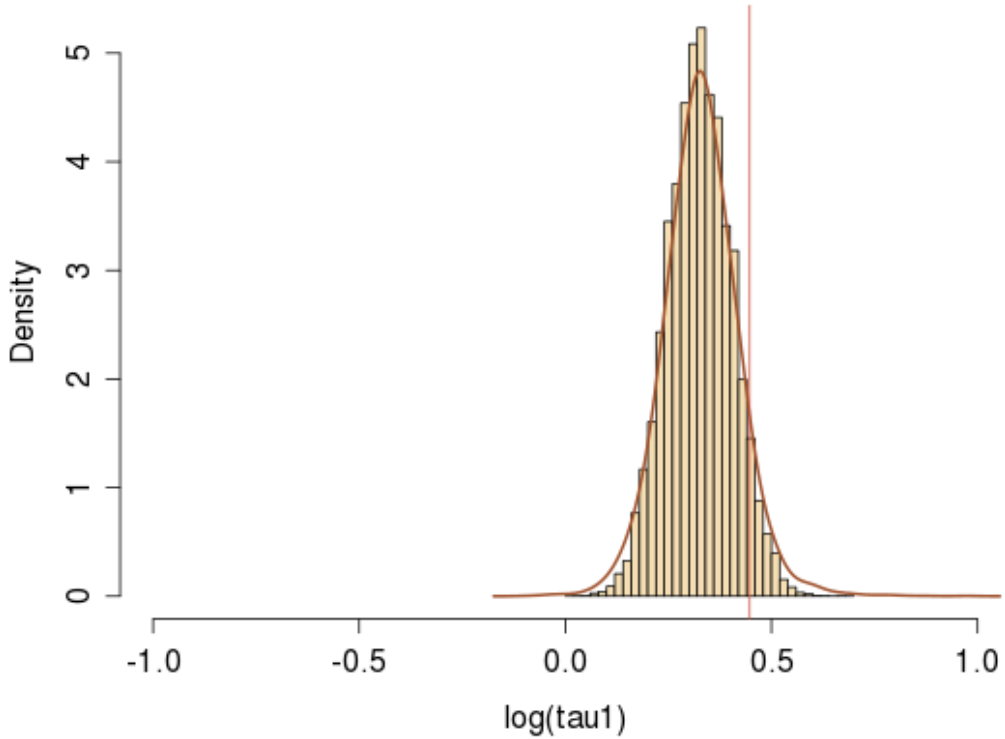
Dataset:

- 50 genes per populations,
- 100 microsat. loci

Assumptions:

- N_e identical over all populations
- $\phi = (\log_{10}\theta, \log_{10}\tau_1, \log_{10}\tau_2)$
- non-informative prior

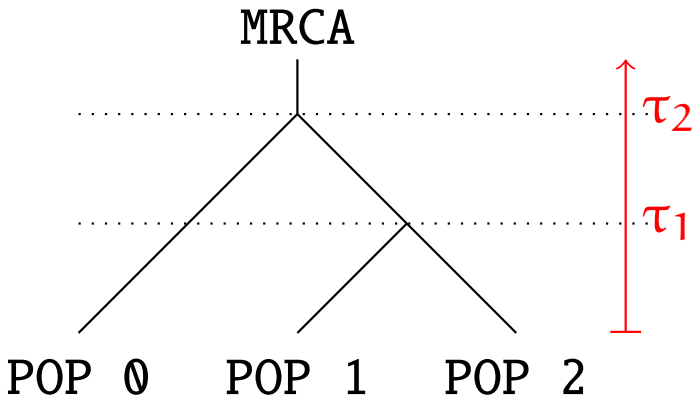
Comparison of the original ABC with ABC_{el}



histogram = ABC_{el}
curve = original ABC
vertical line = "true" parameter

Second experiment

Evolutionary scenario:



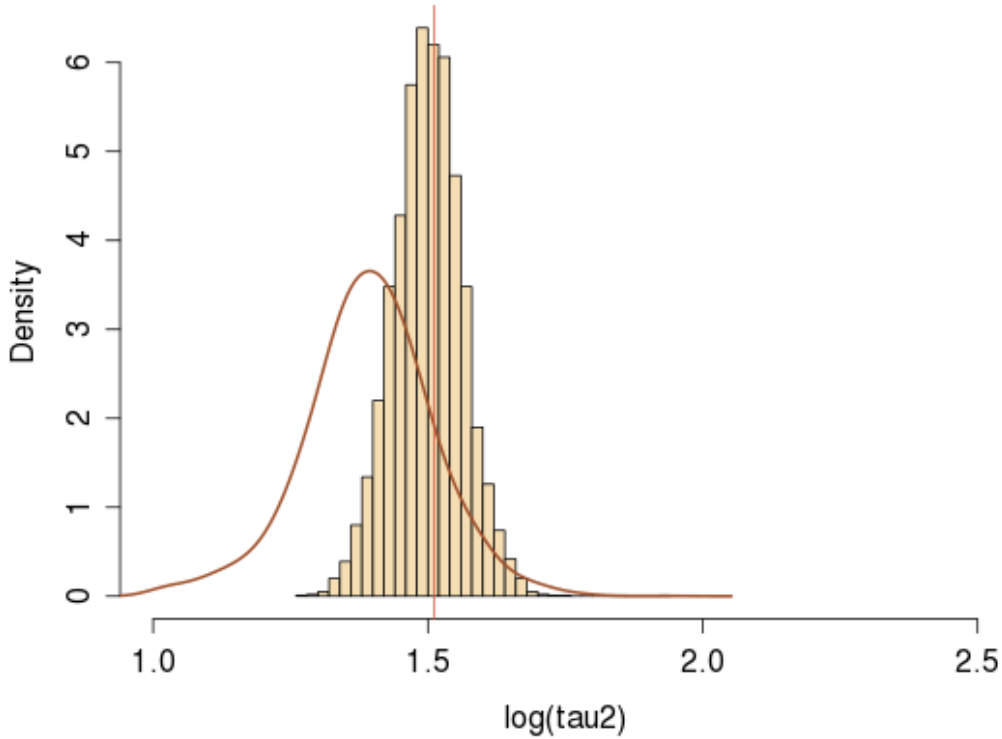
Dataset:

- 50 genes per populations,
- 100 microsat. loci

Assumptions:

- N_e identical over all populations
- $\phi = (\log_{10}\theta, \log_{10}\tau_1, \log_{10}\tau_2)$
- non-informative prior

Comparison of the original ABC with ABC_{el}



histogram = ABC_{el}
curve = original ABC
vertical line = "true" parameter

ABC vs. ABC_{el} on 100 replicates of the 2nd experiment

Accuracy:

	$\log_{10}\theta$		$\log_{10}\tau_1$		$\log_{10}\tau_2$	
	ABC	ABC_{el}	ABC	ABC_{el}	ABC	ABC_{el}
(1)	0.0059	0.0794	0.472	0.483	29.3	4.76
(3)	0.79	0.76	0.88	0.76	0.89	0.79

(1) Root Mean Square Error of the posterior mean

(2) Coverage of the credibility interval of probability 0.8

Computation time: on a recent 6-core computer (C++/OpenMP)

- ABC ≈ 6 hours
- $ABC_{el} \approx 8$ minutes

Why?

On large datasets, ABC_{el} gives more accurate results than ABC

ABC simplifies the dataset through summary statistics
Due to the large dimension of x, the original ABC algorithm estimates

$$\pi\left(\theta \,\middle|\, \eta(x_{obs})\right),$$

where $\eta(x_{obs})$ is some (non-linear) projection of the observed dataset on a space with smaller dimension
$\hookrightarrow$ Some information is lost

ABC_{el} simplifies the model through a generalized moment condition model.
$\hookrightarrow$ Provides more accurate approximation if the constraint is well choosen.

Joint work with

- Christian P. Robert (U. Dauphine & IUF)
- Kerrie Mengersen (QUT, Australia)
- Raphaël Leblois (INRA CBGP, Montpellier)



Grant from ANR through
Project *"Emile"*

First preprint on **arXiv**
Approximate Bayesian computation via empirical likelihood

Coming soon: population genetic models which are too slow to simulate

Joint work with

- Christian P. Robert (U. Dauphine & IUF)
- Kerrie Mengersen (QUT, Australia)
- Raphaël Leblois (INRA CBGP, Montpellier)



Grant from ANR through
Project *"Emile"*

First preprint on **arXiv**
Approximate Bayesian computation via empirical likelihood

Coming soon: population genetic models which are too slow to simulate