Approximate Bayesian Computation: the perspective of a french statistician

Michael G.B. Blum

Laboratoire TIMC, CNRS, Grenoble

CIRM, Marseille, May 2009

Reference

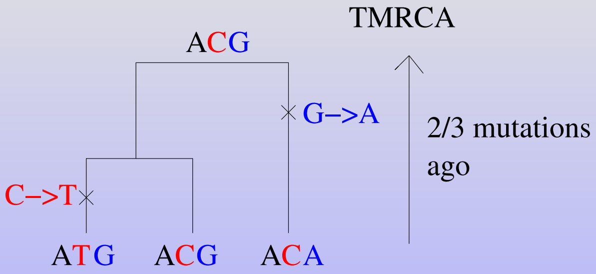
B. Approximate Bayesian computation : a non-parametric perspective. Arxiv :0904 :0635

What is ABC ?

A method of inference well-suited to models for which the likelihood is intractable

What is ABC ?

example : TMRCA (Tavaré et al. 1997; Fu and Li 1997)



What is ABC ?

A simple rejection algorithm

TMRCA example continued

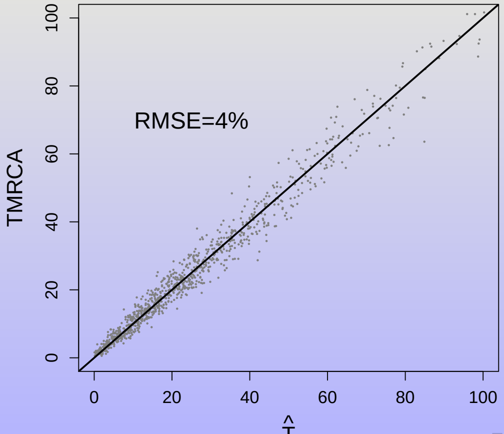
- Simulate the mutation rate θ according to the prior distribution
- Simulate neutral coalescent trees
- Superimpose the mutations according to a Poisson process of rate $\theta/2$

- Accept the coalescent trees for which the simulated genetic diversity $\mathbf{s}_i$ is close enough to the observed ones $\mathbf{s}_{obs}$

$$||\mathbf{s}_i - \mathbf{s}_{obs}|| < \epsilon$$

Does ABC works ?

TMRCA example continued



Does ABC works ?

Does ABC works ?

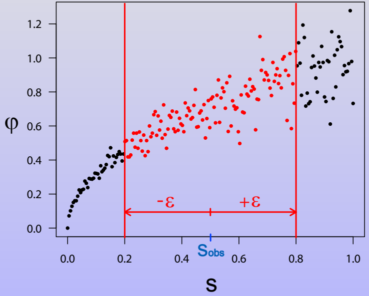
First approximation in ABC

The posterior distribution $P(\Phi|\mathbf{D})$ is replaced by the partial posterior distribution

$$p(\Phi|\mathbf{s}_{obs}) = \frac{p(\mathbf{s}_{obs}|\Phi)p(\Phi)}{p(\mathbf{s}_{obs})}$$

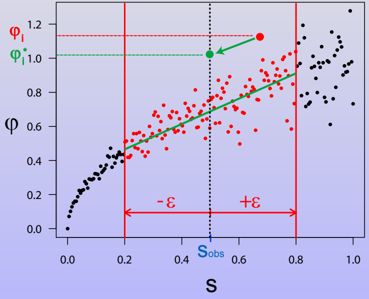
Second approximation in ABC
Estimating the partial posterior distribution

Method 1 : Rejection



Second approximation in ABC
Estimating the partial posterior distribution

Method 2 : Regression adjustment



Second approximation in ABC
Estimating the partial posterior distribution

Method 2 : Regression adjustment

- Local linear regression

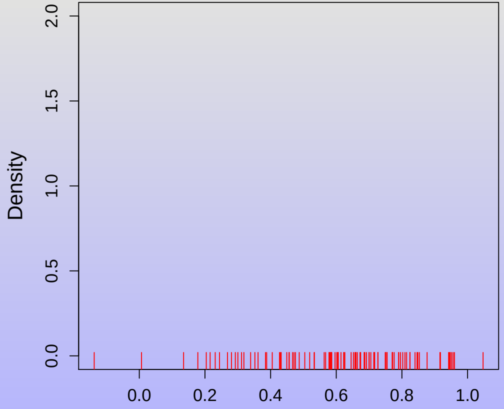
$$\phi_i | \mathbf{s}_i = m(\mathbf{s}_i) + W_i$$

- Adjustement

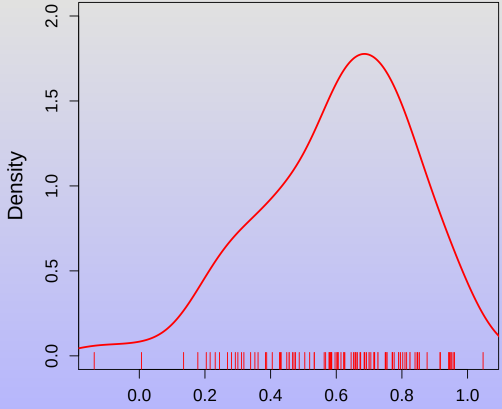
$$\phi_i^* = \hat{m}(\mathbf{s}_{obs}) + \tilde{W}_i,$$

in which the $\tilde{W}_i$'s are the empirical residuals

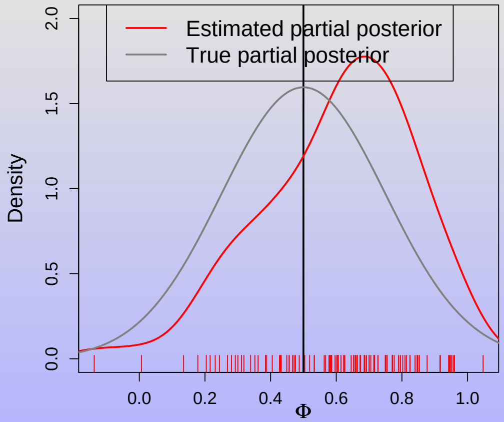
Measuring the error arising from the estimation of the partial posterior distribution



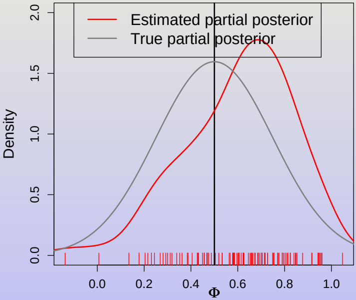
Measuring the error arising from the estimation of the partial posterior distribution



Measuring the error arising from the estimation of the partial posterior distribution



Measuring the error arising from the estimation of the partial posterior distribution



$$MSE = E[(\hat{p}(\phi|\mathbf{s}) - p(\phi|\mathbf{s}))^2], \quad \phi \in \mathbb{R}$$

Main theorem

B. 2009

Asymptotic bias of $\hat{p}(\phi|\mathbf{s}_{obs})$

$$C\epsilon^2$$

Asymptotic variance of $\hat{p}(\phi|\mathbf{s}_{obs})$

$$\frac{C'}{n\epsilon^d}$$

where d is the dimension of the summary statistics and n is the number of simulations.

Conseq 1 : The curse of dimensionality

Effective local size

$$n\epsilon^d$$

To maintain the order of the variance constant, we have

$$\epsilon \propto \left(\frac{1}{n}\right)^{1/d}$$

where d is the dimension of the summary statistics and n is the number of simulations.

Conseq 2 : Difference between the estimators with and without adjustment

Bias for the estimator with quadratic adjustemnt

$$o(\epsilon^2),$$

when the model

$$\phi_i = m(\mathbf{s}_i) + W_i$$

is homoscedastic in the vicinity of $\mathbf{s}_{obs}$.

How many simulations are required to reach a given level of accuracy in ABC

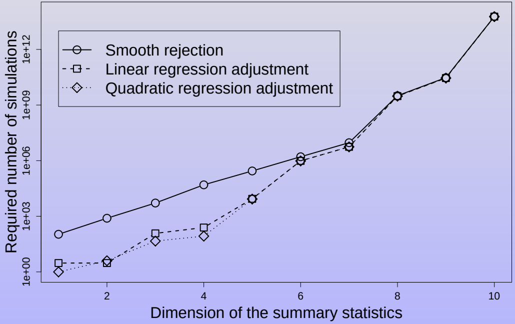
- A standard Gaussian model $(x_1, \ldots, x_d) \rightsquigarrow \mathcal{N}((\mu_1, \ldots, \mu_d), I_d)$.
- Given a sample of $M = 10$ individuals, we can compute, for $\mu_1 = 0$ the asymptotic mean square error (MSE) arising from the estimation of the partial posterior distribution of e^{μ_1}

$$\mathrm{MSE}(n) = \mathrm{bias}^2 + \mathrm{variance}$$

- How many simulations are required so that the relative mean square error is less than 10%

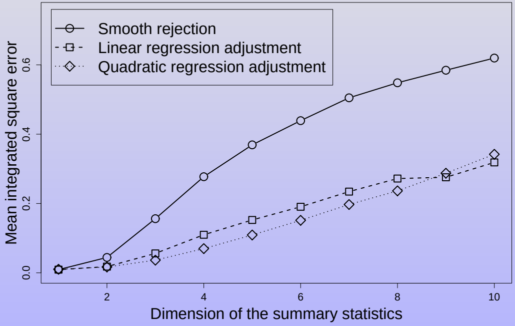
How many simulations are required to reach a given level of accuracy ... continued

The curse of dimensionality



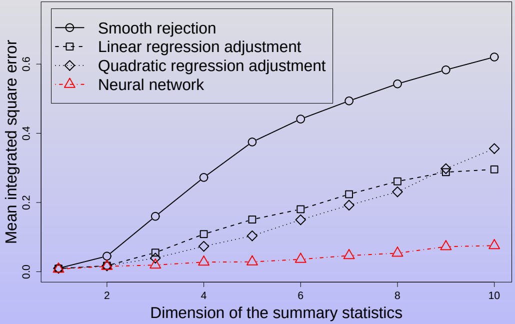
The integrated mean square error as a function of the dimension of the summary statistics d

The curse of dimensionality...continued



Reducing the dimension

Joyce and Marjoram 2008, Blum and François 2009



Reducing the dimension

Blum and François 2009