From model uncertainty to ABC

Christian P. Robert

Université Paris-Dauphine, University of Warwick, & IuF
bayesianstatistics@gmail.com

BIPM Workshop on Measurement Uncertainty, Paris
June 12, 2015

Outline

Introductory notions

The ABC of [Bayesian] statistics

In a classical (Fisher, 1921) perspective, a statistical model is defined by the law of the observations, also called likelihood

$$L(\theta|y_1,\ldots,y_n) = L(\theta|\mathbf{y}) \overset{\text{e.g.}}{=} \prod_{i=1}^{n} f(y_i|\theta)$$

Parameters θ are estimated based on this function $L(\theta|\mathbf{y})$ and on the probabilistic properties of the distribution of the data.

The ABC of [Bayesian] statistics

In a classical (Fisher, 1921) perspective, a statistical model is defined by the law of the observations, also called likelihood

$$L(\theta|y_1,\ldots,y_n) = L(\theta|\mathbf{y}) \overset{\text{e.g.}}{=} \prod_{i=1}^{n} f(y_i|\theta)$$

Parameters θ are estimated based on this function $L(\theta|\mathbf{y})$ and on the probabilistic properties of the distribution of the data.
Comparison of models via likelihoods requires penalization and asymptotics

The ABC of Bayesian statistics

In the Bayesian approach (Bayes, 1763; Laplace, 1773), the parameter is endowed with a probability distribution as well, called the prior distribution and the likelihood becomes a conditional distribution of the data given the parameter, understood as a random variable.

The ABC of Bayesian statistics

In the Bayesian approach (Bayes, 1763; Laplace, 1773), the parameter is endowed with a probability distribution as well, called the prior distribution and the likelihood becomes a conditional distribution of the data given the parameter, understood as a random variable.
Inference based on the posterior distribution, with density

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta)\, L(\theta|\mathbf{y}) \qquad \text{Bayes' Theorem}$$

(also called the posterior)

The ABC of Bayesian statistics

In the Bayesian approach (Bayes, 1763; Laplace, 1773), the parameter is endowed with a probability distribution as well, called the prior distribution and the likelihood becomes a conditional distribution of the data given the parameter, understood as a random variable.

Inference based on the posterior distribution, with density

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta)\, L(\theta|\mathbf{y}) \qquad \text{Bayes' Theorem}$$

(also called the posterior) and model comparison on marginal likelihood

$$m(\mathbf{y}) = \int \pi(\theta)\, L(\theta|\mathbf{y})\, \mathrm{d}\theta$$

A few more details

- The parameter θ does not become a random variable (instead of an unknown constant) in the Bayesian paradygm. Probability calculus is used to quantify the uncertainty about θ as a calibrated quantity.

A few more details

- The parameter θ does not become a random variable (instead of an unknown constant) in the Bayesian paradygm. Probability calculus is used to quantify the uncertainty about θ as a calibrated quantity.
- The prior density $\pi(\cdot)$ is to be understood as a reference measure which, in informative situations, may summarise the available prior information.

A few more details

- The parameter θ does not become a random variable (instead of an unknown constant) in the Bayesian paradygm. Probability calculus is used to quantify the uncertainty about θ as a calibrated quantity.
- The prior density $\pi(\cdot)$ is to be understood as a reference measure which, in informative situations, may summarise the available prior information.
- The impact of the prior density $\pi(\cdot)$ on the resulting inference is real but (mostly) vanishes when the number of observations grows. The only exception is the area of hypothesis testing where both approaches remain unreconcilable.

Getting approximative

Case of a well-defined statistical model where the likelihood function

$$L(\theta|\mathbf{y}) = f(y_1, \ldots, y_n|\theta)$$

is out of reach

Empirical Approximation to the original Bayesian problem

- Degrading the data precision down to tolerance level ε

Getting approximative

Case of a well-defined statistical model where the likelihood function

$$L(\theta|\mathbf{y}) = f(y_1, \ldots, y_n|\theta)$$

is out of reach

Empirical Approximation to the original Bayesian problem

- Degrading the data precision down to tolerance level ε
- Replacing the likelihood with a non-parametric approximation

Getting approximative

Case of a well-defined statistical model where the likelihood function

$$L(\theta|\mathbf{y}) = f(y_1, \ldots, y_n|\theta)$$

is out of reach

Empirical Approximation to the original Bayesian problem

- Degrading the data precision down to tolerance level ε
- Replacing the likelihood with a non-parametric approximation
- Summarising/replacing the data with insufficient statistics

[Marin & al., 2011]

Approximate Bayesian computation

Regular Bayesian computation issues

When faced with a non-standard posterior distribution

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta)L(\theta|\mathbf{y})$$

the standard solution is to use simulation (Monte Carlo) to produce a sample

$$\theta_1, \ldots, \theta_T$$

from $\pi(\theta|\mathbf{y})$ (or approximately by Markov chain Monte Carlo methods)

[Robert & Casella, 2004]

Untractable likelihoods

Cases when the likelihood function $f(\mathbf{y}|\theta)$ is unavailable and when the completion step

$$f(\mathbf{y}|\theta) = \int_{\mathscr{Z}} f(\mathbf{y}, \mathbf{z}|\theta) \, \mathrm{d}\mathbf{z}$$

is impossible or too costly because of the dimension of $\mathbf{z}$
© MCMC cannot be implemented!

Illustrations

Example

Stochastic volatility model: for $t = 1, \ldots, T$,

$$y_t = \exp(z_t)\epsilon_t \ , \ z_t = a + bz_{t-1} + \sigma\eta_t \ ,$$

T very large makes it difficult to include $\mathbf{z}$ within the simulated parameters

Example

Potts model: if $\mathbf{y}$ takes values on a grid $\mathfrak{Y}$ of size k^n and

$$f(\mathbf{y}|\theta) \propto \exp\left\{\theta \sum_{l\sim i} \mathbb{I}_{y_l=y_i}\right\}$$

where $l{\sim}i$ denotes a neighbourhood relation, even moderately large n prohibit the computation of the normalising constant

$$Z_{\boldsymbol{\theta}} = \sum_{\mathbf{y}\in\mathcal{X}} \exp\{\theta S(\mathbf{y})\}$$

with too many terms and poor numerical approximations

[Cucala& al., 2009]

Example

Dynamic mixture model

$$(1 - w_{\mu,\tau}(x))f_{\beta,\lambda}(x) + w_{\mu,\tau}(x)g_{\epsilon,\sigma}(x) \qquad x > 0$$

where $f_{\beta,\lambda}$ is a Weibull density, $g_{\epsilon,\sigma}$ a generalised Pareto density, and $w_{\mu,\tau}$ is the cdf of a Cauchy distribution
Crucially missing the normalising constant

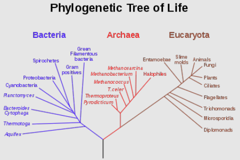
$$\int_0^\infty \{(1 - w_{\mu,\tau}(x))f_{\beta,\lambda}(x) + w_{\mu,\tau}(x)g_{\epsilon,\sigma}(x)\}\,\mathrm{d}x$$

[Frigessi, Haug & Rue, 2002]

Example

Coalescence tree: in population genetics, reconstitution of a common ancestor from a sample of genes via a phylogenetic tree that is close to impossible to integrate out

[100 processor days with 4 parameters]



[Cornuet et al., 2009, Bioinformatics]

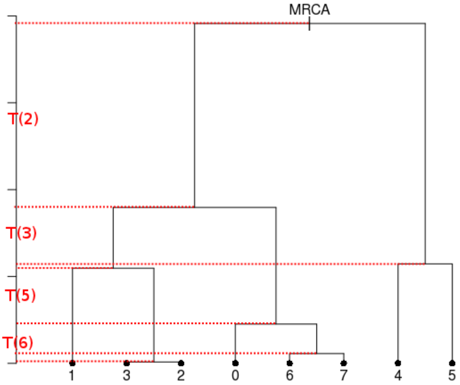
Genetic background of ABC

ABC is a recent computational 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.

[Griffith & al., 1997; Tavaré & al., 1999]

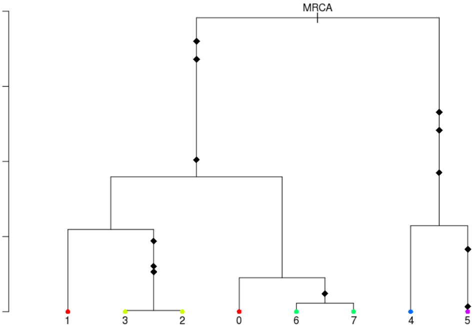
Kingman's coalescent



Kingman's genealogy

When time axis is normalized, $T(k) \sim \text{Exp}(k(k-1)/2)$

Kingman's coalescent



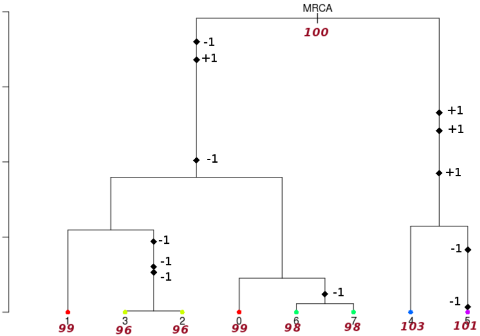
Kingman's genealogy

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

Kingman's coalescent



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)

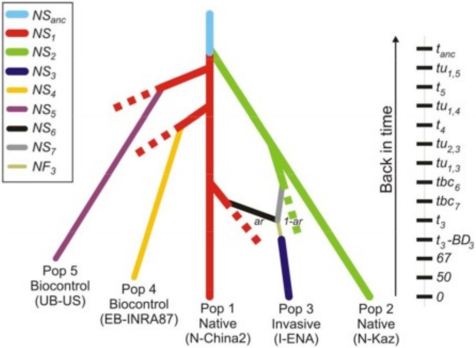
- date of the mutations $\sim$ Poisson process with intensity $\theta/2$ over the branches
- MRCA $= 100$
- independent mutations: ± 1 with pr. $1/2$

Instance of ecological question

- How did the Asian Ladybird beetle arrive in Europe?
- Why do they swarm right now?
- What are the routes of invasion?
- How to get rid of them?

[Lombaert & al., 2010, PLoS ONE]

Worldwide invasion routes of *Harmonia Axyridis*



[Estoup et al., 2012, Molecular Ecology Res.]

© Intractable likelihood

Missing (too much missing!) data structure:

$$f(\mathbf{y}|\boldsymbol{\theta}) = \int_{\mathcal{G}} f(\mathbf{y}|G, \boldsymbol{\theta}) f(G|\boldsymbol{\theta}) \mathrm{d}G$$

cannot be computed in a manageable way...

[Stephens & Donnelly, 2000]

The genealogies are considered as nuisance parameters

Similar exploration of simulation-based and approximation techniques in Econometrics

- Simulated method of moments
- Method of simulated moments
- Simulated pseudo-maximum-likelihood
- Indirect inference

[Gouriéroux & Monfort, 1996]

Econom'ections

Similar exploration of simulation-based and approximation techniques in Econometrics

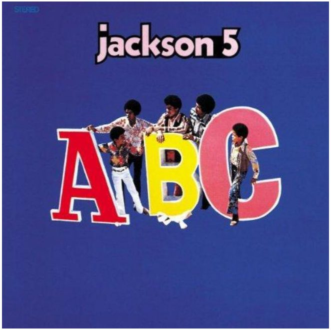
- Simulated method of moments
- Method of simulated moments
- Simulated pseudo-maximum-likelihood
- Indirect inference

[Gouriéroux & Monfort, 1996]

even though motivation is partly-defined models rather than complex likelihoods

A?B?C?

- A stands for approximate [wrong likelihood / picture]
- B stands for Bayesian
- C stands for computation [producing a parameter sample]



The ABC method

Bayesian setting: target is $\pi(\theta) f(x|\theta)$

Bayesian setting: target is $\pi(\theta)f(x|\theta)$

When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

The ABC method

Bayesian setting: target is $\pi(\theta)f(x|\theta)$
When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

ABC algorithm

For an observation $\mathbf{y} \sim f(\mathbf{y}|\theta)$, under the prior $\pi(\theta)$, keep *jointly* simulating

$$\theta' \sim \pi(\theta)\,, \mathbf{z} \sim f(\mathbf{z}|\theta')\,,$$

until the auxiliary variable $\mathbf{z}$ is equal to the observed value, $\mathbf{z} = \mathbf{y}$.

[Tavaré et al., 1997]

Why does it work?!

The proof is trivial:

$$
\begin{aligned}
f(\theta_i) &\propto \sum_{\mathbf{z}\in\mathcal{D}} \pi(\theta_i) f(\mathbf{z}|\theta_i)\mathbb{I}_{\mathbf{y}}(\mathbf{z}) \\
&\propto \pi(\theta_i) f(\mathbf{y}|\theta_i) \\
&= \pi(\theta_i|\mathbf{y})\,.
\end{aligned}
$$

[Accept–Reject 101]

Earlier occurrence

'Bayesian statistics and Monte Carlo methods are ideally suited to the task of passing many models over one dataset'

[Don Rubin, *Annals of Statistics*, 1984]

Note Rubin (1984) does not promote this algorithm for likelihood-free simulation but frequentist intuition on posterior distributions: parameters from posteriors are more likely to be those that could have generated the data.

A as A...pproximative

When y is a continuous random variable, strict equality $\mathbf{z} = \mathbf{y}$ is replaced with a tolerance zone

$$\varrho(\mathbf{y}, \mathbf{z}) \leq \epsilon$$

where ϱ is a distance

A as A...pproximative

When y is a continuous random variable, strict equality $\mathbf{z} = \mathbf{y}$ is replaced with a tolerance zone

$$\varrho(\mathbf{y}, \mathbf{z}) \leq \epsilon$$

where ϱ is a distance
Output distributed from

$$\pi(\theta)\, P_\theta\{\varrho(\mathbf{y}, \mathbf{z}) < \epsilon\} \overset{\text{def}}{\propto} \pi(\theta|\varrho(\mathbf{y}, \mathbf{z}) < \epsilon)$$

[Pritchard et al., 1999]

ABC algorithm

Algorithm 1 Likelihood-free rejection sampler

```
for i = 1 to N do
  repeat
    generate θ' from the prior distribution π(·)
    generate z from the likelihood f(·|θ')
  until ρ{η(z), η(y)} ≤ ε
  set θ_i = θ'
end for
```

where $\eta(\mathbf{y})$ defines a (maybe in-sufficient) statistic

Output

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

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

where $A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) < \epsilon\}$.

Output

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

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

where $A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) < \epsilon\}$.

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

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

Output

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

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

where $A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) < \epsilon\}$.

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

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

Not so good..!

Pima Indian benchmark

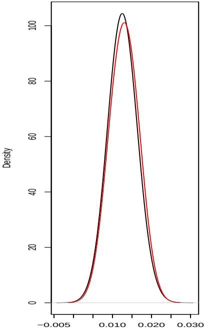


Figure : Comparison between density estimates of the marginals on β_1 (left), β_2 (center) and β_3 (right) from ABC rejection samples (red) and MCMC samples (black)

Which summary?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics [except when done by the experimenters in the field]

Which summary?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics [except when done by the experimenters in the field]

- Loss of statistical information balanced against gain in data roughening
- Approximation error and information loss remain unknown
- Choice of statistics induces choice of distance function towards standardisation

Which summary?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics [except when done by the experimenters in the field]

- Loss of statistical information balanced against gain in data roughening
- Approximation error and information loss remain unknown
- Choice of statistics induces choice of distance function towards standardisation

- may be imposed for external/practical reasons (e.g., DIYABC)
- may gather several non-B point estimates [the more the merrier]
- can [machine-]learn about efficient combination

MA example

Consider the MA(q) model

$$x_t = \epsilon_t + \sum_{i=1}^{q} \vartheta_i \epsilon_{t-i}$$

Simple prior: uniform prior over the identifiability zone, e.g. triangle for MA(2)

MA example (2)

ABC algorithm thus made of

1. picking a new value $(\vartheta_1, \vartheta_2)$ in the triangle
2. generating an iid sequence $(\epsilon_t)_{-q<t\leq T}$
3. producing a simulated series $(x'_t)_{1\leq t\leq T}$

MA example (2)

ABC algorithm thus made of

1. picking a new value $(\vartheta_1, \vartheta_2)$ in the triangle
2. generating an iid sequence $(\epsilon_t)_{-q<t\leq T}$
3. producing a simulated series $(x'_t)_{1\leq t\leq T}$

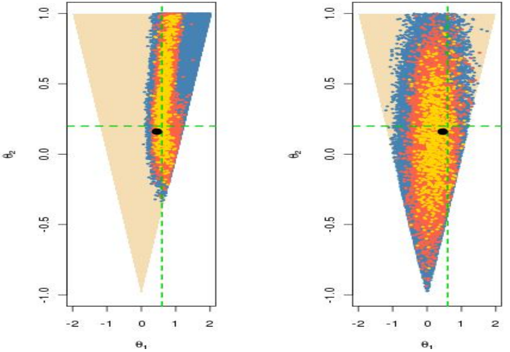
Distance: basic distance between the series

$$\rho((x'_t)_{1\leq t\leq T}, (x_t)_{1\leq t\leq T}) = \sum_{t=1}^{T} (x_t - x'_t)^2$$

or between summary statistics like the first q autocorrelations

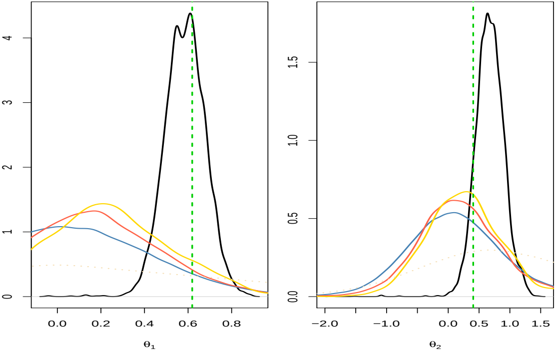
$$\tau_j = \sum_{t=j+1}^{T} x_t x_{t-j}$$

Comparison of distance impact



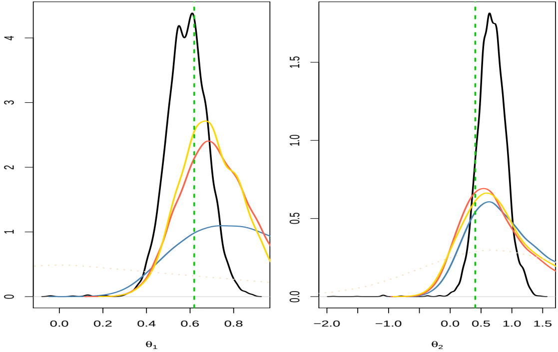
Evaluation of the tolerance on the ABC sample against both distances ($\epsilon = 100\%, 10\%, 1\%, 0.1\%$) for an MA$(2)$ model

Comparison of distance impact



Evaluation of the tolerance on the ABC sample against both distances ($\epsilon = 100\%, 10\%, 1\%, 0.1\%$) for an MA$(2)$ model

Comparison of distance impact



Evaluation of the tolerance on the ABC sample against both distances ($\epsilon = 100\%, 10\%, 1\%, 0.1\%$) for an MA$(2)$ model

ABC advances

Simulating from the prior is often poor in efficiency

ABC advances

Simulating from the prior is often poor in efficiency
Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Beaumont et al., 2009]

[Toni & al., 2009; Fernhead and Prangle, 2012]

ABC advances

Simulating from the prior is often poor in efficiency
Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Beaumont et al., 2009]

[Toni & al., 2009; Fernhead and Prangle, 2012]

...or by viewing the problem as a conditional density estimation and by developing techniques to allow for larger ϵ

[Beaumont et al., 2002; Blum & François, 2010]

ABC advances

Simulating from the prior is often poor in efficiency
Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Beaumont et al., 2009]

[Toni & al., 2009; Fernhead and Prangle, 2012]

...or by viewing the problem as a conditional density estimation and by developing techniques to allow for larger ϵ

[Beaumont et al., 2002; Blum & François, 2010]

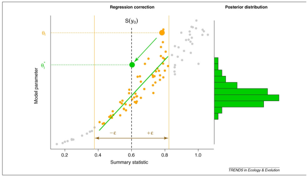
.....or even by including ϵ in the inferential framework [ABC$_\mu$]

[Ratmann et al., 2009]

ABC-NP

Better usage of [prior] simulations by adjustement: instead of throwing away θ' such that $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) > \epsilon$, replace θs with locally regressed

$$\theta^* = \theta - \{\eta(\mathbf{z}) - \eta(\mathbf{y})\}^{\mathsf{T}} \hat{\beta}$$



[Csilléry et al., TEE, 2010]

where $\hat{\beta}$ is obtained by [NP] weighted least square regression on $(\eta(\mathbf{z}) - \eta(\mathbf{y}))$ with weights

$$K_\delta \{\rho(\eta(\mathbf{z}), \eta(\mathbf{y}))\}$$

[Beaumont et al., 2002, Genetics]

attempts at summaries

How to choose the set of summary statistics?

- Joyce and Marjoram (2008, SAGMB)
- Nunes and Balding (2010, SAGMB)
- Fearnhead and Prangle (2012, JRSS B)
- Ratmann et al. (2012, PLOS Comput. Biol)
- Blum et al. (2013, Statistical science)
- EP-ABC of Barthelmé & Chopin (2013, JASA)
- LDA selection of Estoup & al. (2012, Mol. Ecol. Res.)

Semi-automatic ABC

Fearnhead and Prangle (2012) study ABC and selection of summary statistics for parameter estimation

- ABC considered as inferential method and calibrated as such
- randomised (or 'noisy') version of the summary statistics

$$\tilde{\eta}(\mathbf{y}) = \eta(\mathbf{y}) + \tau\epsilon$$

- *optimality* of the posterior expectation

$$\mathbb{E}[\theta|\mathbf{y}]$$

  of the parameter of interest as summary statistics $\eta(\mathbf{y})$!

LDA summaries for model choice

In parallel to F& P semi-automatic ABC, selection of most discriminant subvector out of a collection of summary statistics, can be based on Linear Discriminant Analysis (LDA)

[Estoup & al., 2012, Mol. Ecol. Res.]

Solution now implemented in DIYABC.2

[Cornuet & al., 2008, Bioinf.; Estoup & al., 2013]

LDA advantages

- much faster computation of scenario probabilities via polychotomous regression
- a (much) lower number of explanatory variables improves the accuracy of the ABC approximation, reduces the tolerance ϵ and avoids extra costs in constructing the reference table
- allows for a large collection of initial summaries
- ability to evaluate Type I and Type II errors on more complex models
- LDA reduces correlation among explanatory variables

LDA advantages

- much faster computation of scenario probabilities via polychotomous regression
- a (much) lower number of explanatory variables improves the accuracy of the ABC approximation, reduces the tolerance ϵ and avoids extra costs in constructing the reference table
- allows for a large collection of initial summaries
- ability to evaluate Type I and Type II errors on more complex models
- LDA reduces correlation among explanatory variables

When available, using both simulated and real data sets, posterior probabilities of scenarios computed from LDA-transformed and raw summaries are strongly correlated

BMC Principle

Several models

$$M_1, M_2, \ldots$$

are considered simultaneously for dataset $\mathbf{y}$ and model index $\mathcal{M}$ central to inference.

Use of

- prior $\pi(\mathcal{M} = m)$, plus
- prior distribution on the parameter conditional on the value m of the model index, $\pi_m(\boldsymbol{\theta}_m)$

BMC Principle

Several models

$$M_1, M_2, \dots$$

are considered simultaneously for dataset $\mathbf{y}$ and model index $\mathcal{M}$ central to inference.
Goal is to derive the posterior distribution of $\mathcal{M}$,

$$\pi(\mathcal{M} = m|\text{data})$$

a challenging computational target when models are complex.

Generic ABC for model choice

Algorithm 2 Likelihood-free model choice sampler (ABC-MC)

for $t = 1$ to T **do**
  repeat
    Generate m from the prior $\pi(\mathcal{M} = m)$
    Generate $\boldsymbol{\theta}_m$ from the prior $\pi_m(\boldsymbol{\theta}_m)$
    Generate $\mathbf{z}$ from the model $f_m(\mathbf{z}|\boldsymbol{\theta}_m)$
  until $\rho\{\eta(\mathbf{z}), \eta(\mathbf{y})\} < \epsilon$
  Set $m^{(t)} = m$ and $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_m$
end for

[Grelaud & al., 2009; Toni & al., 2009]

About sufficiency

'Sufficient statistics for individual models are unlikely to be very informative for the model probability.'

[Scott Sisson, Jan. 31, 2011, X.'Og]

About sufficiency

'Sufficient statistics for individual models are unlikely to be very informative for the model probability.'

[Scott Sisson, Jan. 31, 2011, X.'Og]

If $\eta_1(\mathbf{x})$ sufficient statistic for model $m = 1$ and parameter θ_1 and $\eta_2(\mathbf{x})$ sufficient statistic for model $m = 2$ and parameter θ_2, $(\eta_1(\mathbf{x}), \eta_2(\mathbf{x}))$ is not always sufficient for (m, θ_m)

About sufficiency

'Sufficient statistics for individual models are unlikely to be very informative for the model probability.'

[Scott Sisson, Jan. 31, 2011, X.'Og]

If $\eta_1(\mathbf{x})$ sufficient statistic for model $m = 1$ and parameter θ_1 and $\eta_2(\mathbf{x})$ sufficient statistic for model $m = 2$ and parameter θ_2, $(\eta_1(\mathbf{x}), \eta_2(\mathbf{x}))$ is not always sufficient for (m, θ_m)

© **Potential loss of information at the testing level**

Limiting behaviour of B_{12}^{ABC}

When ϵ goes to zero,

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

Limiting behaviour of B_{12}^{ABC}

When ϵ goes to zero,

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

© Bayes factor based on the sole observation of $\eta(\mathbf{y})$

Meaning of the ABC-Bayes factor

'This is also why focus on model discrimination typically (...) proceeds by (...) accepting that the Bayes Factor that one obtains is only derived from the summary statistics and may in no way correspond to that of the full model.'

[Scott Sisson, Jan. 31, 2011, X.'Og]

Meaning of the ABC-Bayes factor

'This is also why focus on model discrimination typically (...) proceeds by (...) accepting that the Bayes Factor that one obtains is only derived from the summary statistics and may in no way correspond to that of the full model.'

[Scott Sisson, Jan. 31, 2011, X.'Og]

In the Poisson/geometric case, if $\mathbb{E}[y_i] = \theta_0 > 0$ and $\eta(\mathbf{y}) = \bar{y}$,

$$\lim_{n\to\infty} B_{12}^{\eta}(\mathbf{y}) = \frac{(\theta_0 + 1)^2}{\theta_0} e^{-\theta_0}$$

The only safe cases

Besides specific models like Gibbs random fields,

using distances over the data itself escapes the discrepancy...

[Toni & Stumpf, 2010;Sousa et al., 2009]

The only safe cases

Besides specific models like Gibbs random fields,

using distances over the data itself escapes the discrepancy...

[Toni & Stumpf, 2010;Sousa et al., 2009]

...but asymptotic consistency of Bayes factors for some summary statistics ensures convergent model choice

[Marin & al., 2014]

Leaning towards machine learning

Main notions:

- ABC-MC seen as learning about which model is most appropriate from a huge (reference) table
- exploiting a large number of summary statistics not an issue for machine learning methods intended to estimate efficient combinations
- abandoning (temporarily?) the idea of estimating posterior probabilities of the models, poorly approximated by machine learning methods, and replacing those by posterior predictive expected loss

Machine learning shift

ABC model choice

A) `Generate a large set of` $(m, \boldsymbol{\theta}, \mathbf{z})$`'s from Bayesian predictive,` $\pi(m)\pi_m(\boldsymbol{\theta})f_m(\mathbf{z}|\boldsymbol{\theta})$

B) `Use machine learning tech. to infer on` $\pi(m|\eta(\mathbf{y}))$

In this perspective:

- (iid) "data set" reference table simulated during stage A)
- observed $\mathbf{y}$ becomes a new data point

Note that:

- predicting m is a classification problem $\iff$ select the best model based on a maximal a posteriori rule, e.g., through **random forests**
- computing $\pi(m|\eta(\mathbf{y}))$ is a regression problem $\iff$ confidence in each model

classification is much simpler than regression (e.g., dim. of objects we try to learn)

Conclusion

- ABC part of a wider picture to handle complex/Big Data models, able to start from rudimentary machine learning summaries
- many formats of empirical [likelihood] Bayes methods available
- lack of comparative tools and of an assessment for information loss
- full Bayesian picture untrustworthy [yet]

Key ideas (for model choice)

- $\pi(m|\eta(\mathbf{y})) \neq \pi(m|\mathbf{y})$
- Rather than approximating $\pi(m|\eta(\mathbf{y}))$, **focus on selecting the best model** (classif. vs regression)
- Assess confidence in the selection with **posterior predictive expected losses**

Conclusion

Key ideas (for model choice)

- $\pi\big(m|\eta(\mathbf{y})\big) \neq \pi\big(m|\mathbf{y}\big)$
- Use **a seasoned machine learning technique** selecting from ABC simulations: minimise 0-1 loss mimics MAP
- Assess confidence in the selection with **posterior predictive expected losses**

Consequences on ABC-PopGen

- Often, **RF $\gg$ k-NN** (less sensible to high correlation in summaries)
- RF requires **many less prior simulations**
- RF selects automatically **relevent summaries**
- **Hence can handle much more complex models**