Approximate Bayesian Computation (ABC): model choice and empirical likelihood

Christian P. Robert

ICERM, “Computational Challenges in Probability”, Nov. 28, 2012

Université Paris-Dauphine, IuF, & CREST
Joint works with J.-M. Cornuet, J.-M. Marin,
K.L. Mengersen, N. Pillai, P. Pudlo and J. Rousseau

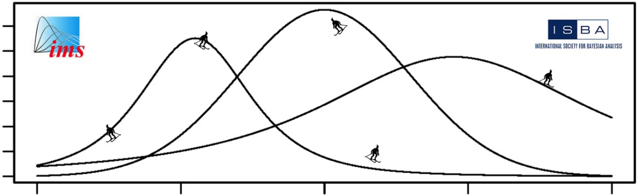
bayesianstatistics@gmail.com

Advertisement

MCMSki IV to be held in Chamonix M^{t} Blanc, France, from Monday, Jan. 6 to Wed., Jan. 8, 2014

All aspects of MCMC^{++} theory and methodology

Parallel (invited and contributed) sessions: call for proposals on website http://www.pages.drexel.edu/ mwl25/mcmski/



Outline

Intractable likelihood

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

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

- is (really!) not available in closed form
- can (easily!) be neither completed nor demarginalised
- cannot be estimated by an unbiased estimator

© Prohibits direct implementation of a generic MCMC algorithm like Metropolis–Hastings

Intractable likelihood

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

$$\ell(\theta|\mathbf{y}) = f(y_1, \dots, y_n|\theta)$$

- is (really!) not available in closed form
- can (easily!) be neither completed nor demarginalised
- cannot be estimated by an unbiased estimator

© Prohibits direct implementation of a generic MCMC algorithm like Metropolis–Hastings

Different perspectives on ABC

What is the (most) fundamental issue?

- a mere computational issue (that will eventually end up being solved by more powerful computers, &tc, even if too costly in the short term)
- an inferential issue (opening opportunities for new inference machine, with different legitimity than classical B approach)
- a Bayesian conundrum (while inferential methods available, how closely related to the B approach?)

Different perspectives on ABC

What is the (most) fundamental issue?

- a mere computational issue (that will eventually end up being solved by more powerful computers, &tc, even if too costly in the short term)
- an inferential issue (opening opportunities for new inference machine, with different legitimity than classical B approach)
- a Bayesian conundrum (while inferential methods available, how closely related to the B approach?)

Different perspectives on ABC

What is the (most) fundamental issue?

- a mere computational issue (that will eventually end up being solved by more powerful computers, &tc, even if too costly in the short term)
- an inferential issue (opening opportunities for new inference machine, with different legitimity than classical B approach)
- a Bayesian conundrum (while inferential methods available, how closely related to the B approach?)

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

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

Indirect inference

Minimise [in θ] a distance between estimators $\hat{\beta}$ based on a pseudo-model for genuine observations and for observations simulated under the true model and the parameter θ.

[Gouriéroux, Monfort, & Renault, 1993; Smith, 1993; Gallant & Tauchen, 1996]

Indirect inference (PML vs. PSE)

Example of the pseudo-maximum-likelihood (PML)

$$\widehat{\beta}(\mathbf{y}) = \arg\max_{\beta} \sum_{t} \log f^{\star}(y_t | \beta, y_{1:(t-1)})$$

leading to

$$\arg\min_{\theta} \|\widehat{\beta}(\mathbf{y}^{o}) - \widehat{\beta}(\mathbf{y}_1(\theta), \ldots, \mathbf{y}_S(\theta))\|^2$$

when

$$\mathbf{y}_s(\theta) \sim f(\mathbf{y}|\theta) \qquad s = 1, \ldots, S$$

Indirect inference (PML vs. PSE)

Example of the pseudo-score-estimator (PSE)

$$\widehat{\beta}(\mathbf{y}) = \arg\min_{\beta} \left\{ \sum_t \frac{\partial \log f^{\star}}{\partial \beta}(y_t|\beta, y_{1:(t-1)}) \right\}^2$$

leading to

$$\arg\min_{\theta} \|\widehat{\beta}(\mathbf{y}^o) - \widehat{\beta}(\mathbf{y}_1(\theta), \dots, \mathbf{y}_S(\theta))\|^2$$

when

$$\mathbf{y}_s(\theta) \sim f(\mathbf{y}|\theta) \qquad s = 1, \dots, S$$

Consistent indirect inference

> *"...in order to get a unique solution the dimension of the auxiliary parameter β must be larger than or equal to the dimension of the initial parameter θ. If the problem is just identified the different methods become easier..."*

Consistency depending on the criterion and on the asymptotic identifiability of θ

[Gouriéroux & Monfort, 1996, p. 66]

Consistent indirect inference

> *"...in order to get a unique solution the dimension of the auxiliary parameter* β *must be larger than or equal to the dimension of the initial parameter* θ*. If the problem is just identified the different methods become easier..."*

Consistency depending on the criterion and on the asymptotic identifiability of θ

[Gouriéroux & Monfort, 1996, p. 66]

Choice of pseudo-model

Arbitrariness of pseudo-model
Pick model such that

1. $\widehat{\beta}(\theta)$ not flat (i.e. sensitive to changes in θ)
2. $\widehat{\beta}(\theta)$ not dispersed (i.e. robust agains changes in $\mathbf{y}^s(\theta)$)

[Frigessi & Heggland, 2004]

Approximate Bayesian computation

Genetic background of ABC

▸ skip genetics

ABC is a recent computational technique that only requires being able to sample from the likelihood $f(\cdot|\theta)$

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

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

Demo-genetic inference

Each model is characterized by a set of parameters θ that cover historical (time divergence, admixture time ...), demographics (population sizes, admixture rates, migration rates, ...) and genetic (mutation rate, ...) factors

The goal is to estimate these parameters from a dataset of polymorphism (DNA sample) $\mathfrak{y}$ observed at the present time

Problem:

most of the time, we cannot calculate the likelihood of the polymorphism data $f(\mathfrak{y}|\theta)$...

Demo-genetic inference

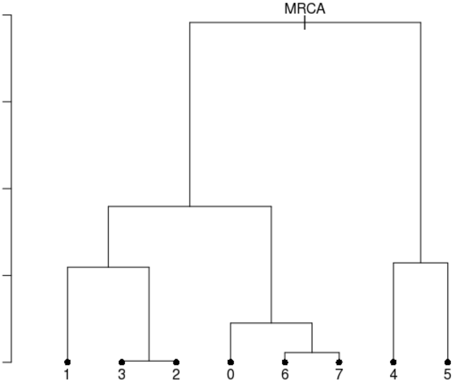
Each model is characterized by a set of parameters θ that cover historical (time divergence, admixture time ...), demographics (population sizes, admixture rates, migration rates, ...) and genetic (mutation rate, ...) factors

The goal is to estimate these parameters from a dataset of polymorphism (DNA sample) $\mathbf{y}$ observed at the present time

Problem:

most of the time, we cannot calculate the likelihood of the polymorphism data $f(\mathbf{y}|\theta)$...

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

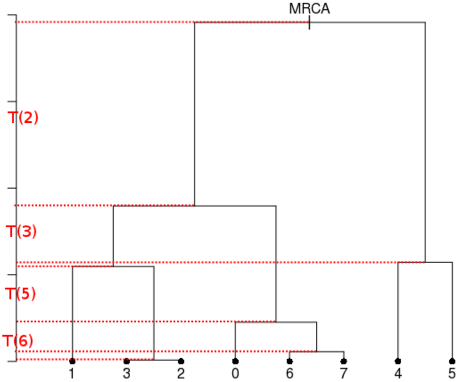


Sample of 8 genes

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. $1/2$

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



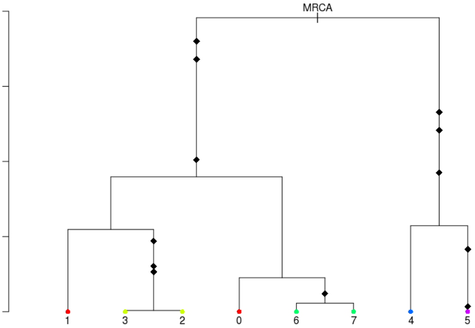
Kingman's genealogy

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
- MRCA $= 100$
- independent mutations: ± 1 with pr. $1/2$

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



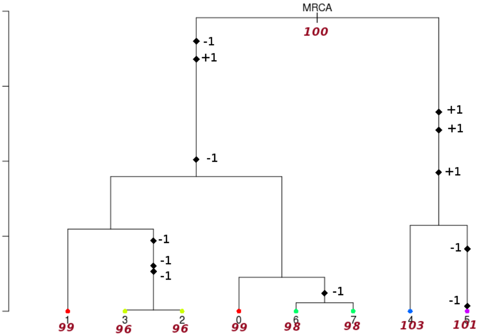
Kingman's genealogy

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

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. $1/2$

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

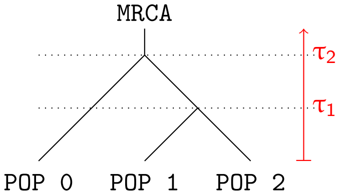
Mutations according to the Simple stepwise Mutation Model (SMM)

- 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 locus**
  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:



Intractable likelihood

Missing (too missing!) data structure:

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

cannot be computed in a manageable way...

The genealogies are considered as nuisance parameters

This modelling clearly differs from the phylogenetic perspective where the tree is the parameter of interest.

Intractable likelihood

Missing (too missing!) data structure:

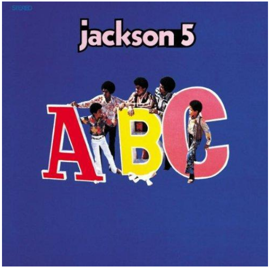
$$f(\mathbf{y}|\theta) = \int_G f(\mathbf{y}|G,\theta) f(G|\theta) \mathrm{d}G$$

cannot be computed in a manageable way...

The genealogies are considered as nuisance parameters

> *This modelling clearly differs from the phylogenetic perspective where the tree is the parameter of interest.*

a dubious ancestry...



You went to school to learn, girl (. . .)
Why 2 plus 2 makes four
Now, now, now, I'm gonna teach you (. . .)

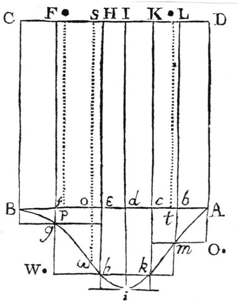
All you gotta do is repeat after me!
A, B, C!
It's easy as 1, 2, 3!
Or simple as Do, Re, Mi! (. . .)

A?B?C?

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

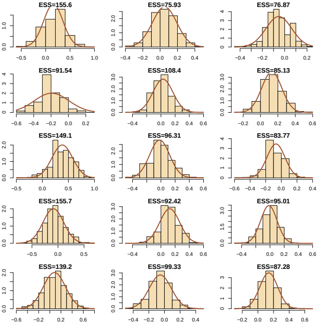
A?B?C?

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



A?B?C?

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



How Bayesian is ABC?

Could we turn the resolution into a Bayesian answer?

- ideally so (not meaningfull: requires ∞-ly powerful computer
- asymptotically so (when sample size goes to ∞: meaningfull?)
- approximation error unknown (w/o costly simulation)
- true Bayes for wrong model (formal and artificial)
- true Bayes for estimated likelihood (back to econometrics?)

Untractable likelihood

Back to stage zero: what can we do when a likelihood function $f(\mathbf{y}|\theta)$ is well-defined but impossible / too costly to compute...?

- MCMC cannot be implemented!
- shall we give up Bayesian inference altogether?!
- or settle for an almost Bayesian inference/picture...?

Untractable likelihood

Back to stage zero: what can we do when a likelihood function $f(\mathbf{y}|\theta)$ is well-defined but impossible / too costly to compute...?

- MCMC cannot be implemented!
- shall we give up Bayesian inference altogether?!
- or settle for an almost Bayesian inference/picture...?

ABC methodology

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

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

Foundation

For an observation $y \sim f(y|\theta)$, under the prior $\pi(\theta)$, if one keeps *jointly* simulating

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

until the auxiliary variable z is equal to the observed value, $z = y$, then the selected

$$\theta' \sim \pi(\theta|y)$$

[Rubin, 1984; Diggle & Gratton, 1984; Tavaré et al., 1997]

ABC methodology

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

Foundation

For an observation $y \sim f(y|\theta)$, under the prior $\pi(\theta)$, if one keeps *jointly* simulating

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

until the auxiliary variable z is equal to the observed value, $z = y$, then the selected

$$\theta' \sim \pi(\theta|y)$$

[Rubin, 1984; Diggle & Gratton, 1984; Tavaré et al., 1997]

ABC methodology

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

Foundation

For an observation $\mathbf{y} \sim f(\mathbf{y}|\theta)$, under the prior $\pi(\theta)$, if one keeps *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, $\mathbf{z} = \mathbf{y}$, then the selected

$$\theta' \sim \pi(\theta|\mathbf{y})$$

[Rubin, 1984; Diggle & Gratton, 1984; Tavaré et al., 1997]

A as A...pproximative

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

$$\rho(\mathbf{y}, \mathbf{z}) \leqslant \epsilon$$

where ρ is a distance
Output distributed from

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

[Pritchard et al., 1999]

A as A...pproximative

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

$$\rho(\mathbf{y}, \mathbf{z}) \leqslant \epsilon$$

where ρ is a distance
Output distributed from

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

[Pritchard et al., 1999]

ABC algorithm

In most implementations, further degree of A...pproximation:

Algorithm 1 Likelihood-free rejection sampler

```
for $i = 1$ to $N$ do
    repeat
        generate $\theta'$ from the prior distribution $\pi(\cdot)$
        generate $z$ from the likelihood $f(\cdot|\theta')$
    until $\rho\{\eta(z), \eta(y)\} \leqslant \epsilon$
    set $\theta_i = \theta'$
end for
```

where $\eta(y)$ defines a (not necessarily sufficient) statistic

Output

The likelihood-free algorithm samples from the marginal in z of:

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

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:

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

...does it?!

Output

The likelihood-free algorithm samples from the marginal in z of:

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

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

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

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

...does it?!

Output

The likelihood-free algorithm samples from the marginal in z of:

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

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

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

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

...does it?!

Output

The likelihood-free algorithm samples from the marginal in 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..!

skip convergence details!

Convergence of ABC

What happens when $\epsilon \to 0$?

For $B \subset \Theta$, we have

$$\begin{aligned}\int_B \frac{\int_{A_{\epsilon,y}} f(z|\theta)\mathrm{d}z}{\int_{A_{\epsilon,y}\times\Theta} \pi(\theta) f(z|\theta)\mathrm{d}z\mathrm{d}\theta}\pi(\theta)\mathrm{d}\theta &= \int_{A_{\epsilon,y}} \frac{\int_B f(z|\theta)\pi(\theta)\mathrm{d}\theta}{\int_{A_{\epsilon,y}\times\Theta} \pi(\theta) f(z|\theta)\mathrm{d}z\mathrm{d}\theta}\mathrm{d}z \\ &= \int_{A_{\epsilon,y}} \frac{\int_B f(z|\theta)\pi(\theta)\mathrm{d}\theta}{m(z)} \frac{m(z)}{\int_{A_{\epsilon,y}\times\Theta} \pi(\theta) f(z|\theta)\mathrm{d}z\mathrm{d}\theta}\mathrm{d}z \\ &= \int_{A_{\epsilon,y}} \pi(B|z)\, \frac{m(z)}{\int_{A_{\epsilon,y}\times\Theta} \pi(\theta) f(z|\theta)\mathrm{d}z\mathrm{d}\theta}\mathrm{d}z\end{aligned}$$

which indicates convergence for a continuous $\pi(B|z)$.

Convergence of ABC

What happens when $\epsilon \to 0$?

For $B \subset \Theta$, we have

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

$$= \int_{A_{\epsilon,\mathbf{y}}} \frac{\int_B f(z|\theta)\pi(\theta)\mathrm{d}\theta}{m(z)} \frac{m(z)}{\int_{A_{\epsilon,\mathbf{y}}\times\Theta} \pi(\theta) f(z|\theta)\mathrm{d}z\mathrm{d}\theta}\mathrm{d}z$$

$$= \int_{A_{\epsilon,\mathbf{y}}} \pi(B|z) \frac{m(z)}{\int_{A_{\epsilon,\mathbf{y}}\times\Theta} \pi(\theta) f(z|\theta)\mathrm{d}z\mathrm{d}\theta}\mathrm{d}z$$

which indicates convergence for a continuous $\pi(B|z)$.

Convergence (do not attempt!)

...and the above does not apply to insufficient statistics:

If $\eta(\mathfrak{y})$ is not a sufficient statistics, the best one can hope for is

$$\pi(\theta|\eta(\mathfrak{y}))\,, \quad \text{not} \quad \pi(\theta|\mathfrak{y})$$

If $\eta(\mathfrak{y})$ is an ancillary statistic, the whole information contained in $\mathfrak{y}$ is lost!, the "best" one can "hope" for is

$$\pi(\theta|\eta(\mathfrak{y})) = \pi(\theta)$$

Bummer!!!

Convergence (do not attempt!)

...and the above does not apply to insufficient statistics:

If $\eta(\mathbf{y})$ is not a sufficient statistics, the best one can hope for is

$$\pi(\theta|\eta(\mathbf{y}))\,, \quad \text{not} \quad \pi(\theta|\mathbf{y})$$

If $\eta(\mathbf{y})$ is an ancillary statistic, the whole information contained in $\mathbf{y}$ is lost!, the "best" one can "hope" for is

$$\pi(\theta|\eta(\mathbf{y})) = \pi(\theta)$$

Bummer!!!

Convergence (do not attempt!)

...and the above does not apply to insufficient statistics:

If $\eta(\mathbf{y})$ is not a sufficient statistics, the best one can hope for is

$$\pi(\theta|\eta(\mathbf{y}))\,, \quad \text{not} \quad \pi(\theta|\mathbf{y})$$

If $\eta(\mathbf{y})$ is an ancillary statistic, the whole information contained in $\mathbf{y}$ is lost!, the "best" one can "hope" for is

$$\pi(\theta|\eta(\mathbf{y})) = \pi(\theta)$$

Bummer!!!

Convergence (do not attempt!)

...and the above does not apply to insufficient statistics:

If $\eta(\mathbf{y})$ is not a sufficient statistics, the best one can hope for is

$$\pi(\theta|\eta(\mathbf{y}))\,, \quad \text{not} \quad \pi(\theta|\mathbf{y})$$

If $\eta(\mathbf{y})$ is an ancillary statistic, the whole information contained in $\mathbf{y}$ is lost!, the "best" one can "hope" for is

$$\pi(\theta|\eta(\mathbf{y})) = \pi(\theta)$$

Bummer!!!

MA example

Inference on the parameters of a MA(q) model

$$x_t = \epsilon_t + \sum_{i=1}^{q} \vartheta_i \epsilon_{t-i} \qquad \epsilon_{t-i} \text{ i.i.d.w.n.}$$

▸ bypass MA illustration

Simple prior: uniform over the inverse [real and complex] roots in

$$\mathcal{Q}(u) = 1 - \sum_{i=1}^{q} \vartheta_i u^i$$

under the identifiability conditions

MA example

Inference on the parameters of a MA(q) model

$$x_t = \epsilon_t + \sum_{i=1}^{q} \vartheta_i \epsilon_{t-i} \qquad \epsilon_{t-i} \text{ i.i.d.w.n.}$$

▸ bypass MA illustration

Simple prior: uniform prior over the identifiability zone in the parameter space, i.e. 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 $(\epsilon_t)_{-q<t\leqslant T}$
3. producing a simulated series $(x'_t)_{1\leqslant t\leqslant T}$

Distance: basic distance between the series

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

or distance between summary statistics like the $q = 2$ autocorrelations

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

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 $(\epsilon_t)_{-q<t\leqslant T}$
3. producing a simulated series $(x'_t)_{1\leqslant t\leqslant T}$

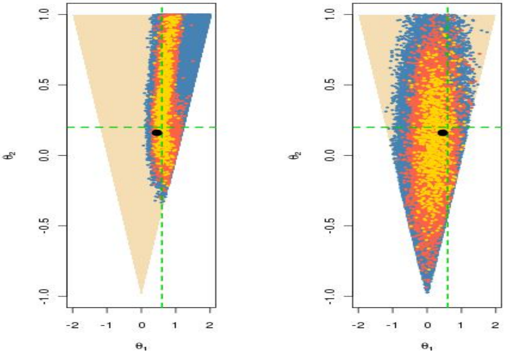
Distance: basic distance between the series

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

or distance between summary statistics like the $q = 2$ autocorrelations

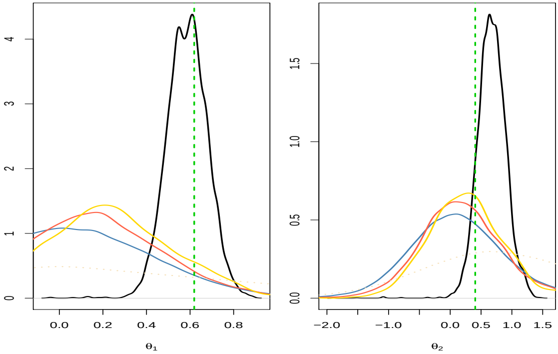
$$\tau_j = \sum_{t=j+1}^{T} x_t x_{t-j}$$

Comparison of distance impact



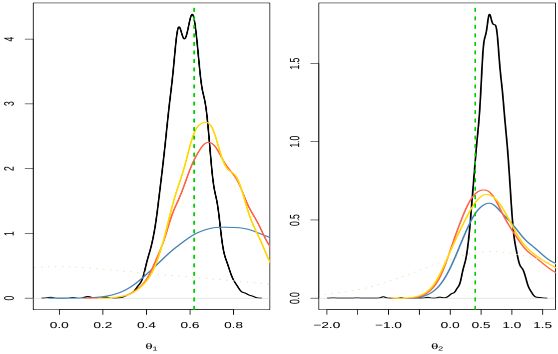
Impact of tolerance on ABC sample against either distance ($\epsilon = 100\%, 10\%, 1\%, 0.1\%$) for an MA$(2)$ model

Comparison of distance impact



Impact of tolerance on ABC sample against either distance ($\epsilon = 100\%, 10\%, 1\%, 0.1\%$) for an MA$(2)$ model

Comparison of distance impact



Impact of tolerance on ABC sample against either distance ($\epsilon = 100\%, 10\%, 1\%, 0.1\%$) for an MA(2) model

Comments

- Role of distance paramount (because $\epsilon \neq 0$)
- Scaling of components of $\eta(\mathbf{y})$ is also determinant
- ϵ matters little if “small enough”
- representative of “curse of dimensionality”
- small is beautiful!
- the data as a whole may be paradoxically weakly informative for ABC

ABC (simul') advances

▸ how approximative is ABC? ▸ ABC as knn

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; Bortot et al., 2007, Sisson et al., 2007]

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

[Beaumont et al., 2002]

.....or even by including ϵ in the inferential framework [ABC_μ]

[Ratmann et al., 2009]

ABC (simul') advances

▸ how approximative is ABC? ▸ ABC as knn

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; Bortot et al., 2007, Sisson et al., 2007]

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

[Beaumont et al., 2002]

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

[Ratmann et al., 2009]

ABC (simul') advances

▸ how approximative is ABC?

▸ ABC as knn

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; Bortot et al., 2007, Sisson et al., 2007]

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

[Beaumont et al., 2002]

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

[Ratmann et al., 2009]

ABC (simul') advances

▸ how approximative is ABC?

▸ ABC as knn

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; Bortot et al., 2007, Sisson et al., 2007]

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

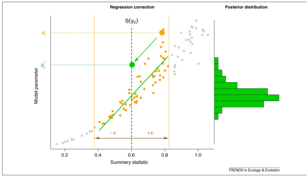
[Beaumont et al., 2002]

.....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 θ's with locally regressed transforms



[Csilléry et al., TEE, 2010]

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

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]

ABC-NP (regression)

Also found in the subsequent literature, e.g. in ▸ Fearnhead-Prangle (2012): weight directly simulation by

$$K_\delta\left\{\rho(\eta(\boldsymbol{z}(\theta)),\eta(\boldsymbol{y}))\right\}$$

or

$$\frac{1}{S}\sum_{s=1}^{S} K_\delta\left\{\rho(\eta(\boldsymbol{z}^s(\theta)),\eta(\boldsymbol{y}))\right\}$$

[consistent estimate of $f(\eta|\theta)$]

Curse of dimensionality: poor estimate when $d = \dim(\eta)$ is large...

ABC-NP (regression)

Also found in the subsequent literature, e.g. in ▸ Fearnhead-Prangle (2012): weight directly simulation by

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

or

$$\frac{1}{S}\sum_{s=1}^{S} K_\delta\{\rho(\eta(\mathbf{z}^s(\theta)), \eta(\mathbf{y}))\}$$

[consistent estimate of $f(\eta|\theta)$]

Curse of dimensionality: poor estimate when $d = \dim(\eta)$ is large...

ABC-NP (density estimation)

Use of the kernel weights

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

leads to the NP estimate of the posterior expectation

$$\frac{\sum_i \theta_i K_\delta \left\{\rho(\eta(\mathbf{z}(\theta_i)), \eta(\mathbf{y}))\right\}}{\sum_i K_\delta \left\{\rho(\eta(\mathbf{z}(\theta_i)), \eta(\mathbf{y}))\right\}}$$

[Blum, JASA, 2010]

ABC-NP (density estimation)

Use of the kernel weights

$$K_\delta\left\{\rho(\eta(\boldsymbol{z}(\theta)),\eta(\mathbf{y}))\right\}$$

leads to the NP estimate of the posterior conditional density

$$\frac{\sum_i \tilde{K}_b(\theta_i-\theta)K_\delta\left\{\rho(\eta(\boldsymbol{z}(\theta_i)),\eta(\mathbf{y}))\right\}}{\sum_i K_\delta\left\{\rho(\eta(\boldsymbol{z}(\theta_i)),\eta(\mathbf{y}))\right\}}$$

[Blum, JASA, 2010]

ABC-NP (density estimations)

Other versions incorporating regression adjustments

$$\frac{\sum_i \tilde{K}_b(\theta_i^* - \theta) K_\delta \{\rho(\eta(\boldsymbol{z}(\theta_i)), \eta(\mathbf{y}))\}}{\sum_i K_\delta \{\rho(\eta(\boldsymbol{z}(\theta_i)), \eta(\mathbf{y}))\}}$$

In all cases, error

$$\mathbb{E}[\hat{g}(\theta|\mathbf{y})] - g(\theta|\mathbf{y}) = cb^2 + c\delta^2 + \mathrm{O}_P(b^2 + \delta^2) + \mathrm{O}_P(1/n\delta^d)$$

$$\mathrm{var}(\hat{g}(\theta|\mathbf{y})) = \frac{c}{nb\delta^d}(1 + \mathrm{o}_P(1))$$

ABC-NP (density estimations)

Other versions incorporating regression adjustments

$$\frac{\sum_i \tilde{K}_b(\theta_i^* - \theta) K_\delta \{\rho(\eta(\mathbf{z}(\theta_i)), \eta(\mathbf{y}))\}}{\sum_i K_\delta \{\rho(\eta(\mathbf{z}(\theta_i)), \eta(\mathbf{y}))\}}$$

In all cases, error

$$\begin{aligned}\mathbb{E}[\hat{g}(\theta|\mathbf{y})] - g(\theta|\mathbf{y}) &= cb^2 + c\delta^2 + \mathrm{O}_P(b^2 + \delta^2) + \mathrm{O}_P(1/n\delta^d) \\ \mathrm{var}(\hat{g}(\theta|\mathbf{y})) &= \frac{c}{nb\delta^d}(1 + \mathrm{o}_P(1))\end{aligned}$$

[Blum, JASA, 2010]

ABC-NP (density estimations)

Other versions incorporating regression adjustments

$$\frac{\sum_i \tilde{K}_b(\theta_i^* - \theta) K_\delta \{\rho(\eta(\mathbf{z}(\theta_i)), \eta(\mathbf{y}))\}}{\sum_i K_\delta \{\rho(\eta(\mathbf{z}(\theta_i)), \eta(\mathbf{y}))\}}$$

In all cases, error

$$\begin{aligned}\mathbb{E}[\hat{g}(\theta|\mathbf{y})] - g(\theta|\mathbf{y}) &= cb^2 + c\delta^2 + O_P(b^2 + \delta^2) + O_P(1/n\delta^d) \\ \mathrm{var}(\hat{g}(\theta|\mathbf{y})) &= \frac{c}{nb\delta^d}(1 + o_P(1))\end{aligned}$$

[standard NP calculations]

ABC-NCH

Incorporating non-linearities and heterocedasticities:

$$\theta^* = \hat{m}(\eta(\mathbf{y})) + [\theta - \hat{m}(\eta(\mathbf{z}))] \frac{\hat{\sigma}(\eta(\mathbf{y}))}{\hat{\sigma}(\eta(\mathbf{z}))}$$

where

- $\hat{m}(\eta)$ estimated by non-linear regression (e.g., neural network)
- $\hat{\sigma}(\eta)$ estimated by non-linear regression on residuals

$$\log\{\theta_i - \hat{m}(\eta_i)\}^2 = \log \sigma^2(\eta_i) + \xi_i$$

[Blum & François, 2009]

ABC-NCH

Incorporating non-linearities and heterocedasticities:

$$\theta^* = \hat{m}(\eta(\mathbf{y})) + [\theta - \hat{m}(\eta(\mathbf{z}))] \frac{\hat{\sigma}(\eta(\mathbf{y}))}{\hat{\sigma}(\eta(\mathbf{z}))}$$

where

- $\hat{m}(\eta)$ estimated by non-linear regression (e.g., neural network)
- $\hat{\sigma}(\eta)$ estimated by non-linear regression on residuals

$$\log\{\theta_i - \hat{m}(\eta_i)\}^2 = \log \sigma^2(\eta_i) + \xi_i$$

[Blum & François, 2009]

ABC as knn

[Biau et al., 2012, arxiv:1207.6461]

Practice of ABC: determine tolerance ϵ as a quantile on observed distances, say 10% or 1% quantile,

$$\epsilon = \epsilon_N = q_\alpha(d_1, \ldots, d_N)$$

- Interpretation of ε as nonparametric bandwidth only approximation of the actual practice

  [Blum & François, 2010]

- ABC is a k-nearest neighbour (knn) method with $k_N = N\epsilon_N$

  [Loftsgaarden & Quesenberry, 1965]

ABC as knn

[Biau et al., 2012, arxiv:1207.6461]

Practice of ABC: determine tolerance ϵ as a quantile on observed distances, say 10% or 1% quantile,

$$\epsilon = \epsilon_N = q_\alpha(d_1, \dots, d_N)$$

- Interpretation of ε as nonparametric bandwidth only approximation of the actual practice
[Blum & François, 2010]
- ABC is a k-nearest neighbour (knn) method with $k_N = N\epsilon_N$
[Loftsgaarden & Quesenberry, 1965]

ABC as knn

[Biau et al., 2012, arxiv:1207.6461]

Practice of ABC: determine tolerance ϵ as a quantile on observed distances, say 10% or 1% quantile,

$$\epsilon = \epsilon_N = q_\alpha(d_1, \ldots, d_N)$$

- Interpretation of ε as nonparametric bandwidth only approximation of the actual practice
  [Blum & François, 2010]
- ABC is a k-nearest neighbour (knn) method with $k_N = N\epsilon_N$
  [Loftsgaarden & Quesenberry, 1965]

ABC consistency

Provided

$$k_N/\log\log N \longrightarrow \infty \quad \text{and} \quad k_N/N \longrightarrow 0$$

as $N \to \infty$, for almost all s_0 (with respect to the distribution of S), with probability 1,

$$\frac{1}{k_N}\sum_{j=1}^{k_N} \varphi(\theta_j) \longrightarrow \mathbb{E}[\varphi(\theta_j)|S=s_0]$$

[Devroye, 1982]

Biau et al. (2012) also recall pointwise and integrated mean square error consistency results on the corresponding kernel estimate of the conditional posterior distribution, under constraints

$$k_N \to \infty, \quad k_N/N \to 0, \quad h_N \to 0 \quad \text{and} \quad h_N^p k_N \to \infty,$$

ABC consistency

Provided

$$k_N / \log\log N \longrightarrow \infty \quad \text{and} \quad k_N/N \longrightarrow 0$$

as $N \to \infty$, for almost all s_0 (with respect to the distribution of S), with probability 1,

$$\frac{1}{k_N} \sum_{j=1}^{k_N} \varphi(\theta_j) \longrightarrow \mathbb{E}[\varphi(\theta_j)|S = s_0]$$

[Devroye, 1982]

Biau et al. (2012) also recall pointwise and integrated mean square error consistency results on the corresponding kernel estimate of the conditional posterior distribution, under constraints

$$k_N \to \infty, \quad k_N/N \to 0, \quad h_N \to 0 \quad \text{and} \quad h_N^p k_N \to \infty,$$

Rates of convergence

Further assumptions (on target and kernel) allow for precise (integrated mean square) convergence rates (as a power of the sample size N), derived from classical k-nearest neighbour regression, like

- when $m = 1, 2, 3$, $k_N \approx N^{(p+4)/(p+8)}$ and rate $N^{-\frac{4}{p+8}}$
- when $m = 4$, $k_N \approx N^{(p+4)/(p+8)}$ and rate $N^{-\frac{4}{p+8}} \log N$
- when $m > 4$, $k_N \approx N^{(p+4)/(m+p+4)}$ and rate $N^{-\frac{4}{m+p+4}}$

[Biau et al., 2012, arxiv:1207.6461]

Only applies to sufficient summary statistics

Rates of convergence

Further assumptions (on target and kernel) allow for precise (integrated mean square) convergence rates (as a power of the sample size N), derived from classical k-nearest neighbour regression, like

- when $m = 1, 2, 3$, $k_N \approx N^{(p+4)/(p+8)}$ and rate $N^{-\frac{4}{p+8}}$
- when $m = 4$, $k_N \approx N^{(p+4)/(p+8)}$ and rate $N^{-\frac{4}{p+8}} \log N$
- when $m > 4$, $k_N \approx N^{(p+4)/(m+p+4)}$ and rate $N^{-\frac{4}{m+p+4}}$

[Biau et al., 2012, arxiv:1207.6461]

Only applies to sufficient summary statistics

ABC inference machine

Introduction

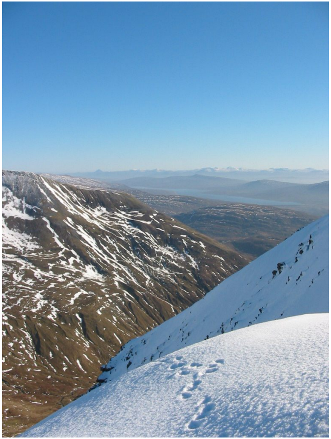
ABC

ABC as an inference machine
- Error inc.
- Exact BC and approximate targets
- summary statistic

ABC for model choice

Model choice consistency

$\mathrm{ABC}_{\mathrm{el}}$



How much Bayesian ABC is..?

- maybe a convergent method of inference (meaningful? sufficient? foreign?)
- approximation error unknown (w/o simulation)
- pragmatic Bayes (there is no other solution!)
- many calibration issues (tolerance, distance, statistics)

How much Bayesian ABC is..?

- maybe a convergent method of inference (meaningful? sufficient? foreign?)
- approximation error unknown (w/o simulation)
- pragmatic Bayes (there is no other solution!)
- many calibration issues (tolerance, distance, statistics)

...should Bayesians care?!

How much Bayesian ABC is..?

- maybe a convergent method of inference (meaningful? sufficient? foreign?)
- approximation error unknown (w/o simulation)
- pragmatic Bayes (there is no other solution!)
- many calibration issues (tolerance, distance, statistics)

yes they should!!!

How much Bayesian ABC is..?

- maybe a convergent method of inference (meaningful? sufficient? foreign?)
- approximation error unknown (w/o simulation)
- pragmatic Bayes (there is no other solution!)
- many calibration issues (tolerance, distance, statistics)

to ABC_{el}

ABC$_\mu$

Idea Infer about the error as well as about the parameter:
Use of a joint density

$$f(\theta, \epsilon|y) \propto \xi(\epsilon|y, \theta) \times \pi_\theta(\theta) \times \pi_\epsilon(\epsilon)$$

where y is the data, and $\xi(\epsilon|y, \theta)$ is the prior predictive density of $\rho(\eta(z), \eta(y))$ given θ and y when $z \sim f(z|\theta)$

Warning! Replacement of $\xi(\epsilon|y, \theta)$ with a non-parametric kernel approximation.

[Ratmann, Andrieu, Wiuf and Richardson, 2009, PNAS]

ABC$_\mu$

Idea Infer about the error as well as about the parameter:
Use of a joint density

$$f(\theta,\epsilon|\mathbf{y}) \propto \xi(\epsilon|\mathbf{y},\theta) \times \pi_\theta(\theta) \times \pi_\epsilon(\epsilon)$$

where $\mathbf{y}$ is the data, and $\xi(\epsilon|\mathbf{y},\theta)$ is the prior predictive density of $\rho(\eta(\mathbf{z}),\eta(\mathbf{y}))$ given θ and $\mathbf{y}$ when $\mathbf{z} \sim f(\mathbf{z}|\theta)$
Warning! Replacement of $\xi(\epsilon|\mathbf{y},\theta)$ with a non-parametric kernel approximation.

[Ratmann, Andrieu, Wiuf and Richardson, 2009, PNAS]

ABC$_\mu$

Idea Infer about the error as well as about the parameter:
Use of a joint density

$$f(\theta, \epsilon|\mathbf{y}) \propto \xi(\epsilon|\mathbf{y}, \theta) \times \pi_\theta(\theta) \times \pi_\epsilon(\epsilon)$$

where $\mathbf{y}$ is the data, and $\xi(\epsilon|\mathbf{y}, \theta)$ is the prior predictive density of $\rho(\eta(\mathbf{z}), \eta(\mathbf{y}))$ given θ and $\mathbf{y}$ when $\mathbf{z} \sim f(\mathbf{z}|\theta)$
Warning! Replacement of $\xi(\epsilon|\mathbf{y}, \theta)$ with a non-parametric kernel approximation.

[Ratmann, Andrieu, Wiuf and Richardson, 2009, PNAS]

ABC$_\mu$ details

Multidimensional distances ρ_k $(k = 1, \ldots, K)$ and errors $\epsilon_k = \rho_k(\eta_k(\mathbf{z}), \eta_k(\mathbf{y}))$, with

$$\epsilon_k \sim \xi_k(\epsilon|\mathbf{y}, \theta) \approx \hat{\xi}_k(\epsilon|\mathbf{y}, \theta) = \frac{1}{Bh_k} \sum_b K[\{\epsilon_k - \rho_k(\eta_k(\mathbf{z}_b), \eta_k(\mathbf{y}))\}/h_k]$$

then used in replacing $\xi(\epsilon|\mathbf{y}, \theta)$ with $\min_k \hat{\xi}_k(\epsilon|\mathbf{y}, \theta)$

ABC$_\mu$ involves acceptance probability

$$\frac{\pi(\theta', \epsilon')}{\pi(\theta, \epsilon)} \frac{q(\theta', \theta) q(\epsilon', \epsilon)}{q(\theta, \theta') q(\epsilon, \epsilon')} \frac{\min_k \hat{\xi}_k(\epsilon'|\mathbf{y}, \theta')}{\min_k \hat{\xi}_k(\epsilon|\mathbf{y}, \theta)}$$

ABC$_\mu$ details

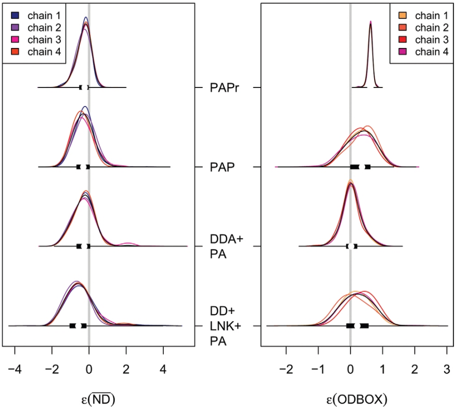
Multidimensional distances ρ_k $(k = 1, \ldots, K)$ and errors $\epsilon_k = \rho_k(\eta_k(\mathbf{z}), \eta_k(\mathbf{y}))$, with

$$\epsilon_k \sim \xi_k(\epsilon|\mathbf{y}, \theta) \approx \hat{\xi}_k(\epsilon|\mathbf{y}, \theta) = \frac{1}{Bh_k} \sum_b K[\{\epsilon_k - \rho_k(\eta_k(\mathbf{z}_b), \eta_k(\mathbf{y}))\}/h_k]$$

then used in replacing $\xi(\epsilon|\mathbf{y}, \theta)$ with $\min_k \hat{\xi}_k(\epsilon|\mathbf{y}, \theta)$
ABC$_\mu$ involves acceptance probability

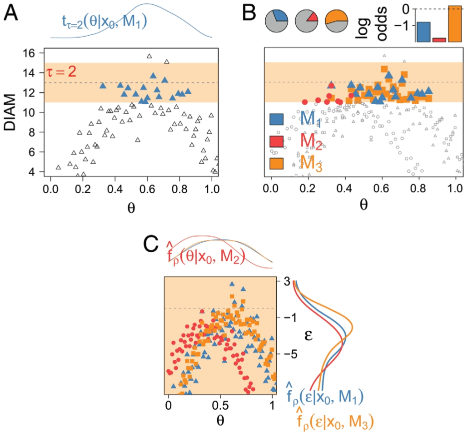
$$\frac{\pi(\theta', \epsilon')}{\pi(\theta, \epsilon)} \frac{q(\theta', \theta) q(\epsilon', \epsilon)}{q(\theta, \theta') q(\epsilon, \epsilon')} \frac{\min_k \hat{\xi}_k(\epsilon'|\mathbf{y}, \theta')}{\min_k \hat{\xi}_k(\epsilon|\mathbf{y}, \theta)}$$

ABC_{μ} multiple errors



[© Ratmann et al., PNAS, 2009]

ABC_μ for model choice



[© Ratmann et al., PNAS, 2009]

Wilkinson's exact BC (not exactly!)

ABC approximation error (i.e. non-zero tolerance) replaced with exact simulation from a controlled approximation to the target, convolution of true posterior with kernel function

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

with K_ϵ kernel parameterised by bandwidth ϵ.

[Wilkinson, 2008]

Theorem

The ABC algorithm based on the assumption of a randomised observation $\mathbf{y} = \tilde{\mathbf{y}} + \xi$, $\xi \sim K_\epsilon$, and an acceptance probability of

$$K_\epsilon(\mathbf{y} - \mathbf{z})/M$$

gives draws from the posterior distribution $\pi(\theta|\mathbf{y})$.

Wilkinson's exact BC (not exactly!)

ABC approximation error (i.e. non-zero tolerance) replaced with exact simulation from a controlled approximation to the target, convolution of true posterior with kernel function

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

with K_ϵ kernel parameterised by bandwidth ϵ.

[Wilkinson, 2008]

Theorem

The ABC algorithm based on the assumption of a randomised observation $\mathbf{y} = \tilde{\mathbf{y}} + \xi$, $\xi \sim K_\epsilon$, and an acceptance probability of

$$K_\epsilon(\mathbf{y}-\mathbf{z})/M$$

gives draws from the posterior distribution $\pi(\theta|\mathbf{y})$.

How exact a BC?

> *"Using ϵ to represent measurement error is straightforward, whereas using ϵ to model the model discrepancy is harder to conceptualize and not as commonly used"*
>
> [Richard Wilkinson, 2008]

How exact a BC?

Pros

- Pseudo-data from *true* model and observed data from *noisy* model
- Interesting perspective in that outcome is completely controlled
- Link with ◂ ABC$_\mu$ and assuming $\mathfrak{y}$ is observed with a measurement error with density K_ϵ
- Relates to the theory of model approximation

[Kennedy & O'Hagan, 2001]

Cons

- Requires K_ϵ to be bounded by M
- True approximation error never assessed
- Requires a modification of the standard ABC algorithm

ABC for HMMs

Specific case of a hidden ▸ Markov model

$$\begin{aligned} X_{t+1} &\sim Q_\theta(X_t, \cdot) \\ Y_{t+1} &\sim g_\theta(\cdot | x_t) \end{aligned}$$

where only $\mathfrak{y}^0_{1:n}$ is observed.

[Dean, Singh, Jasra, & Peters, 2011]

Use of specific constraints, adapted to the Markov structure:

$$\left\{y_1 \in \mathcal{B}(y_1^0, \epsilon)\right\} \times \cdots \times \left\{y_n \in \mathcal{B}(y_n^0, \epsilon)\right\}$$

ABC for HMMs

Specific case of a hidden ▸ Markov model

$$\begin{aligned} X_{t+1} &\sim Q_\theta(X_t, \cdot) \\ Y_{t+1} &\sim g_\theta(\cdot|x_t) \end{aligned}$$

where only $\mathfrak{y}^0_{1:n}$ is observed.

[Dean, Singh, Jasra, & Peters, 2011]

Use of specific constraints, adapted to the Markov structure:

$$\left\{y_1 \in \mathcal{B}(y_1^0, \epsilon)\right\} \times \cdots \times \left\{y_n \in \mathcal{B}(y_n^0, \epsilon)\right\}$$

ABC-MLE for HMMs

ABC-MLE defined by

$$\hat{\theta}_n^\epsilon = \arg\max_\theta \mathbb{P}_\theta \left(Y_1 \in \mathcal{B}(y_1^0, \epsilon), \ldots, Y_n \in \mathcal{B}(y_n^0, \epsilon) \right)$$

Exact MLE for the likelihood ▸ same basis as Wilkinson!

$$p_\theta^\epsilon(y_1^0, \ldots, y_n)$$

corresponding to the perturbed process

$$(x_t, y_t + \epsilon z_t)_{1 \leqslant t \leqslant n} \qquad z_t \sim \mathcal{U}(\mathcal{B}(0,1)$$

[Dean, Singh, Jasra, & Peters, 2011]

ABC-MLE for HMMs

ABC-MLE defined by

$$\hat{\theta}_n^\epsilon = \arg\max_\theta \mathbb{P}_\theta \left(Y_1 \in \mathcal{B}(y_1^0, \epsilon), \ldots, Y_n \in \mathcal{B}(y_n^0, \epsilon) \right)$$

Exact MLE for the likelihood ▸ same basis as Wilkinson!

$$p_\theta^\epsilon(y_1^0, \ldots, y_n)$$

corresponding to the perturbed process

$$(x_t, y_t + \epsilon z_t)_{1 \leqslant t \leqslant n} \quad z_t \sim \mathcal{U}(\mathcal{B}(0,1)$$

[Dean, Singh, Jasra, & Peters, 2011]

ABC-MLE is biased

- ABC-MLE is asymptotically (in n) biased with target

$$l^{\epsilon}(\theta) = \mathbb{E}_{\theta^*}[\log p_{\theta}^{\epsilon}(Y_1|Y_{-\infty:0})]$$

- but ABC-MLE converges to true value in the sense

$$l^{\epsilon_n}(\theta_n) \to l^{\epsilon}(\theta)$$

for all sequences (θ_n) converging to θ and $\epsilon_n \searrow \epsilon$

ABC-MLE is biased

- ABC-MLE is asymptotically (in n) biased with target

$$l^{\epsilon}(\theta) = \mathbb{E}_{\theta^*}[\log p_{\theta}^{\epsilon}(Y_1 | Y_{-\infty:0})]$$

- but ABC-MLE converges to true value in the sense

$$l^{\epsilon_n}(\theta_n) \rightarrow l^{\epsilon}(\theta)$$

for all sequences (θ_n) converging to θ and $\epsilon_n \searrow \epsilon$

Noisy ABC-MLE

Idea: Modify instead the data from the start

$$(y_1^0 + \epsilon\zeta_1, \dots, y_n + \epsilon\zeta_n)$$

[▸ see Fearnhead-Prangle]

noisy ABC-MLE estimate

$$\arg\max_\theta \mathbb{P}_\theta\left(Y_1 \in \mathcal{B}(y_1^0 + \epsilon\zeta_1, \epsilon), \dots, Y_n \in \mathcal{B}(y_n^0 + \epsilon\zeta_n, \epsilon)\right)$$

[Dean, Singh, Jasra, & Peters, 2011]

Consistent noisy ABC-MLE

- Degrading the data improves the estimation performances:
  - Noisy ABC-MLE is asymptotically (in n) consistent
  - under further assumptions, the noisy ABC-MLE is asymptotically normal
  - increase in variance of order ϵ^{-2}
- likely degradation in precision or computing time due to the lack of summary statistic [curse of dimensionality]

SMC for ABC likelihood

Algorithm 2 SMC ABC for HMMs

```
Given $\theta$
for $k = 1, \ldots, n$ do
    generate proposals $(x_k^1, y_k^1), \ldots, (x_k^N, y_k^N)$ from the model
    weigh each proposal with $\omega_k^l = \mathbb{I}_{\mathcal{B}(y_k^0 + \epsilon\zeta_k, \epsilon)}(y_k^l)$
    renormalise the weights and sample the $x_k^l$'s accordingly
end for
approximate the likelihood by
$$\prod_{k=1}^{n} \left( \sum_{l=1}^{N} \omega_k^l / N \right)$$
```

[Jasra, Singh, Martin, & McCoy, 2010]

Which summary?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics

Starting from a large collection of summary statistics is available, Joyce and Marjoram (2008) consider the sequential inclusion into the ABC target, with a stopping rule based on a likelihood ratio test

- Not taking into account the sequential nature of the tests
- Depends on parameterisation
- Order of inclusion matters
- likelihood ratio test?!

Which summary?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics
Starting from a large collection of summary statistics is available, Joyce and Marjoram (2008) consider the sequential inclusion into the ABC target, with a stopping rule based on a likelihood ratio test

- Not taking into account the sequential nature of the tests
- Depends on parameterisation
- Order of inclusion matters
- likelihood ratio test?!

Which summary?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics
Starting from a large collection of summary statistics is available, Joyce and Marjoram (2008) consider the sequential inclusion into the ABC target, with a stopping rule based on a likelihood ratio test

- Not taking into account the sequential nature of the tests
- Depends on parameterisation
- Order of inclusion matters
- likelihood ratio test?!

Which summary for model choice?

Depending on the choice of $\eta(\cdot)$, the Bayes factor based on this insufficient statistic,

$$B^{\eta}_{12}(\mathbf{y}) = \frac{\int \pi_1(\theta_1) f^{\eta}_1(\eta(\mathbf{y})|\theta_1)\, d\theta_1}{\int \pi_2(\theta_2) f^{\eta}_2(\eta(\mathbf{y})|\theta_2)\, d\theta_2}\,,$$

is consistent or not.

[X, Cornuet, Marin, & Pillai, 2012]

Consistency only depends on the range of $\mathbb{E}_i[\eta(\mathbf{y})]$ under both models.

[Marin, Pillai, X, & Rousseau, 2012]

Which summary for model choice?

Depending on the choice of $\eta(\cdot)$, the Bayes factor based on this insufficient statistic,

$$B_{12}^{\eta}(\mathbf{y}) = \frac{\int \pi_1(\theta_1) f_1^{\eta}(\eta(\mathbf{y})|\theta_1)\,\mathrm{d}\theta_1}{\int \pi_2(\theta_2) f_2^{\eta}(\eta(\mathbf{y})|\theta_2)\,\mathrm{d}\theta_2}\,,$$

is consistent or not.

[X, Cornuet, Marin, & Pillai, 2012]

Consistency only depends on the range of $\mathbb{E}_i[\eta(\mathbf{y})]$ under both models.

[Marin, Pillai, X, & Rousseau, 2012]

Semi-automatic ABC

Fearnhead and Prangle (2010) study ABC and the selection of the summary statistic in close proximity to ◂ Wilkinson's proposal

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

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

- derivation of a well-calibrated version of ABC, i.e. an algorithm that gives proper predictions for the distribution associated with this randomised summary statistic

Summary [of F&P/statistics)

- optimality of the posterior expectation

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

  of the parameter of interest as summary statistics $\eta(\mathbf{y})$!
- use of the standard quadratic loss function

$$(\theta-\theta_0)^{\mathrm{T}}A(\theta-\theta_0)\,.$$

- recent extension to model choice, optimality of Bayes factor

$$B_{12}(\mathbf{y})$$

[F&P, ISBA 2012 talk]

Summary [of F&P/statistics)

- optimality of the posterior expectation

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

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

- use of the standard quadratic loss function

$$(\theta - \theta_0)^{\mathsf{T}} A (\theta - \theta_0)\,.$$

- recent extension to model choice, optimality of Bayes factor

$$B_{12}(\mathbf{y})$$

[F&P, ISBA 2012 talk]

Conclusion

- Choice of summary statistics is paramount for ABC validation/performance
- At best, ABC approximates $\pi(. \mid \eta(y))$
- Model selection feasible with ABC [with caution!]
- For estimation, consistency if $\{\theta; \mu(\theta) = \mu_0\} = \theta_0$
- For testing consistency if $\{\mu_1(\theta_1), \theta_1 \in \Theta_1\} \cap \{\mu_2(\theta_2), \theta_2 \in \Theta_2\} = \emptyset$

[Marin et al., 2011]

Conclusion

- Choice of summary statistics is paramount for ABC validation/performance
- At best, ABC approximates $\pi(. \mid \eta(\mathbf{y}))$
- Model selection feasible with ABC [with caution!]
- For estimation, consistency if $\{\theta; \mu(\theta) = \mu_0\} = \theta_0$
- For testing consistency if $\{\mu_1(\theta_1), \theta_1 \in \Theta_1\} \cap \{\mu_2(\theta_2), \theta_2 \in \Theta_2\} = \emptyset$

[Marin et al., 2011]

Conclusion

- Choice of summary statistics is paramount for ABC validation/performance
- At best, ABC approximates $\pi(. \mid \eta(\mathbf{y}))$
- Model selection feasible with ABC [with caution!]
- For estimation, consistency if $\{\theta; \mu(\theta) = \mu_0\} = \theta_0$
- For testing consistency if $\{\mu_1(\theta_1), \theta_1 \in \Theta_1\} \cap \{\mu_2(\theta_2), \theta_2 \in \Theta_2\} = \emptyset$

[Marin et al., 2011]

Conclusion

- Choice of summary statistics is paramount for ABC validation/performance
- At best, ABC approximates $\pi(. \mid \eta(\mathbf{y}))$
- Model selection feasible with ABC [with caution!]
- For estimation, consistency if $\{\theta; \mu(\theta) = \mu_0\} = \theta_0$
- For testing consistency if $\{\mu_1(\theta_1), \theta_1 \in \Theta_1\} \cap \{\mu_2(\theta_2), \theta_2 \in \Theta_2\} = \emptyset$

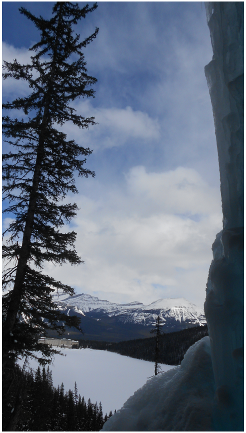
[Marin et al., 2011]

Conclusion

- Choice of summary statistics is paramount for ABC validation/performance
- At best, ABC approximates $\pi(.\mid\eta(\mathbf{y}))$
- Model selection feasible with ABC [with caution!]
- For estimation, consistency if $\{\theta;\mu(\theta)=\mu_0\}=\theta_0$
- For testing consistency if $\{\mu_1(\theta_1),\theta_1\in\Theta_1\}\cap\{\mu_2(\theta_2),\theta_2\in\Theta_2\}=\emptyset$

[Marin et al., 2011]

ABC for model choice



Bayesian model choice

BMC Principle

Several models

$$M_1, M_2, \ldots$$

are considered simultaneously for dataset $\mathfrak{y}$ 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, $\pi_m(\theta_m)$

Bayesian model choice

BMC Principle

Several models

$$M_1, M_2, \ldots$$

are considered simultaneously for dataset $\mathfrak{y}$ 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 3 Likelihood-free model choice sampler (ABC-MC)

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

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

ABC estimates

Posterior probability $\pi(\mathcal{M} = m|\mathbf{y})$ approximated by the frequency of acceptances from model m

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{I}_{m^{(t)}=m} \,.$$

ABC estimates

Posterior probability $\pi(\mathcal{M} = m|\mathbf{y})$ approximated by the frequency of acceptances from model m

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{I}_{m^{(t)}=m} \,.$$

Extension to a weighted polychotomous logistic regression estimate of $\pi(\mathcal{M} = m|\mathbf{y})$, with non-parametric kernel weights

[Cornuet et al., DIYABC, 2009]

Potts model

▸ Skip MRFs

Potts model

Distribution with an energy function of the form

$$\theta S(\mathbf{y}) = \theta \sum_{l \sim i} \delta_{y_l = y_i}$$

where $l \sim i$ denotes a neighbourhood structure

In most realistic settings, summation

$$Z_\theta = \sum_{x \in \mathcal{X}} \exp\{\theta S(x)\}$$

involves too many terms to be manageable and numerical approximations cannot always be trusted

Potts model

▸ Skip MRFs

Potts model

Distribution with an energy function of the form

$$\theta S(\mathbf{y}) = \theta \sum_{l \sim i} \delta_{y_l = y_i}$$

where $l \sim i$ denotes a neighbourhood structure

In most realistic settings, summation

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

involves too many terms to be manageable and numerical approximations cannot always be trusted

Neighbourhood relations

Setup

Choice to be made between M neighbourhood relations

$$i \overset{m}{\sim} i' \qquad (0 \leqslant m \leqslant M-1)$$

with

$$S_m(\boldsymbol{x}) = \sum_{i \overset{m}{\sim} i'} \mathbb{I}_{\{x_i = x_{i'}\}}$$

driven by the posterior probabilities of the models.

Model index

Computational target:

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) \propto \int_{\Theta_m} f_m(\mathbf{x}|\theta_m)\pi_m(\theta_m)\,\mathrm{d}\theta_m\,\pi(\mathcal{M} = m)$$

If $S(\mathbf{x})$ sufficient statistic for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$,

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) = \mathbb{P}(\mathcal{M} = m|S(\mathbf{x}))\,.$$

Model index

Computational target:

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) \propto \int_{\Theta_m} f_m(\mathbf{x}|\theta_m)\pi_m(\theta_m)\,\mathrm{d}\theta_m\,\pi(\mathcal{M} = m)$$

If $S(\mathbf{x})$ sufficient statistic for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$,

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) = \mathbb{P}(\mathcal{M} = m|S(\mathbf{x}))\,.$$

Sufficient statistics in Gibbs random fields

Each model m has its own sufficient statistic $S_m(\cdot)$ and $S(\cdot) = (S_0(\cdot), \ldots, S_{M-1}(\cdot))$ is also (model-)sufficient.

Explanation: For Gibbs random fields,

$$\begin{aligned} x|\mathcal{M} = m \sim f_m(\mathbf{x}|\theta_m) &= f_m^1(\mathbf{x}|S(\mathbf{x}))f_m^2(S(\mathbf{x})|\theta_m) \\ &= \frac{1}{n(S(\mathbf{x}))} f_m^2(S(\mathbf{x})|\theta_m) \end{aligned}$$

where

$$n(S(\mathbf{x})) = \sharp\{\tilde{\mathbf{x}} \in \mathcal{X} : S(\tilde{\mathbf{x}}) = S(\mathbf{x})\}$$

© $S(\mathbf{x})$ is sufficient for the joint parameters

Sufficient statistics in Gibbs random fields

Each model m has its own sufficient statistic $S_m(\cdot)$ and $S(\cdot) = (S_0(\cdot), \ldots, S_{M-1}(\cdot))$ is also (model-)sufficient.
Explanation: For Gibbs random fields,

$$\begin{aligned} x|\mathcal{M} = m \sim f_m(\mathbf{x}|\theta_m) &= f_m^1(\mathbf{x}|S(\mathbf{x}))f_m^2(S(\mathbf{x})|\theta_m) \\ &= \frac{1}{n(S(\mathbf{x}))} f_m^2(S(\mathbf{x})|\theta_m) \end{aligned}$$

where

$$n(S(\mathbf{x})) = \sharp\{\tilde{\mathbf{x}} \in \mathcal{X} : S(\tilde{\mathbf{x}}) = S(\mathbf{x})\}$$

ⓒ $S(x)$ is sufficient for the joint parameters

Toy example

iid Bernoulli model versus two-state first-order Markov chain, i.e.

$$f_0(\mathbf{x}|\theta_0) = \exp\left(\theta_0 \sum_{i=1}^{n} \mathbb{I}_{\{x_i=1\}}\right) \Big/ \{1+\exp(\theta_0)\}^n \,,$$

versus

$$f_1(\mathbf{x}|\theta_1) = \frac{1}{2} \exp\left(\theta_1 \sum_{i=2}^{n} \mathbb{I}_{\{x_i=x_{i-1}\}}\right) \Big/ \{1+\exp(\theta_1)\}^{n-1} \,,$$

with priors $\theta_0 \sim \mathcal{U}(-5,5)$ and $\theta_1 \sim \mathcal{U}(0,6)$ (inspired by "phase transition" boundaries).

About sufficiency

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

© Potential loss of information at the testing level

About sufficiency

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

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

Poisson/geometric example

Sample

$$\boldsymbol{x} = (x_1, \ldots, x_n)$$

from either a Poisson $\mathcal{P}(\lambda)$ or from a geometric $\mathcal{G}(p)$
Sum

$$S = \sum_{i=1}^{n} x_i = \eta(\boldsymbol{x})$$

sufficient statistic for either model but not simultaneously

Limiting behaviour of B_{12} ($T \to \infty$)

ABC approximation

$$\widehat{B_{12}}(\mathbf{y}) = \frac{\sum_{t=1}^{T} \mathbb{I}_{m^t=1}\, \mathbb{I}_{\rho\{\eta(\mathbf{z}^t),\eta(\mathbf{y})\}\leqslant\epsilon}}{\sum_{t=1}^{T} \mathbb{I}_{m^t=2}\, \mathbb{I}_{\rho\{\eta(\mathbf{z}^t),\eta(\mathbf{y})\}\leqslant\epsilon}} \,,$$

where the (m^t, z^t)'s are simulated from the (joint) prior

As T goes to infinity, limit

$$\begin{aligned} B_{12}^{\epsilon}(\mathbf{y}) &= \frac{\int \mathbb{I}_{\rho\{\eta(\mathbf{z}),\eta(\mathbf{y})\}\leqslant\epsilon} \pi_1(\theta_1) f_1(\mathbf{z}|\theta_1)\, \mathrm{d}\mathbf{z}\, \mathrm{d}\theta_1}{\int \mathbb{I}_{\rho\{\eta(\mathbf{z}),\eta(\mathbf{y})\}\leqslant\epsilon} \pi_2(\theta_2) f_2(\mathbf{z}|\theta_2)\, \mathrm{d}\mathbf{z}\, \mathrm{d}\theta_2} \\ &= \frac{\int \mathbb{I}_{\rho\{\eta,\eta(\mathbf{y})\}\leqslant\epsilon} \pi_1(\theta_1) f_1^{\eta}(\eta|\theta_1)\, \mathrm{d}\eta\, \mathrm{d}\theta_1}{\int \mathbb{I}_{\rho\{\eta,\eta(\mathbf{y})\}\leqslant\epsilon} \pi_2(\theta_2) f_2^{\eta}(\eta|\theta_2)\, \mathrm{d}\eta\, \mathrm{d}\theta_2} \,, \end{aligned}$$

where $f_1^{\eta}(\eta|\theta_1)$ and $f_2^{\eta}(\eta|\theta_2)$ distributions of $\eta(\mathbf{z})$

Limiting behaviour of B_{12} ($T \to \infty$)

ABC approximation

$$\widehat{B_{12}}(\mathbf{y}) = \frac{\sum_{t=1}^{T} \mathbb{I}_{m^t=1}\, \mathbb{I}_{\rho\{\eta(\mathbf{z}^t),\eta(\mathbf{y})\}\leqslant\epsilon}}{\sum_{t=1}^{T} \mathbb{I}_{m^t=2}\, \mathbb{I}_{\rho\{\eta(\mathbf{z}^t),\eta(\mathbf{y})\}\leqslant\epsilon}}\,,$$

where the (m^t, z^t)'s are simulated from the (joint) prior
As T goes to infinity, limit

$$\begin{aligned} B_{12}^{\epsilon}(\mathbf{y}) &= \frac{\int \mathbb{I}_{\rho\{\eta(\mathbf{z}),\eta(\mathbf{y})\}\leqslant\epsilon}\pi_1(\theta_1) f_1(\mathbf{z}|\theta_1)\,\mathrm{d}\mathbf{z}\,\mathrm{d}\theta_1}{\int \mathbb{I}_{\rho\{\eta(\mathbf{z}),\eta(\mathbf{y})\}\leqslant\epsilon}\pi_2(\theta_2) f_2(\mathbf{z}|\theta_2)\,\mathrm{d}\mathbf{z}\,\mathrm{d}\theta_2} \\ &= \frac{\int \mathbb{I}_{\rho\{\eta,\eta(\mathbf{y})\}\leqslant\epsilon}\pi_1(\theta_1) f_1^{\eta}(\eta|\theta_1)\,\mathrm{d}\eta\,\mathrm{d}\theta_1}{\int \mathbb{I}_{\rho\{\eta,\eta(\mathbf{y})\}\leqslant\epsilon}\pi_2(\theta_2) f_2^{\eta}(\eta|\theta_2)\,\mathrm{d}\eta\,\mathrm{d}\theta_2}\,, \end{aligned}$$

where $f_1^{\eta}(\eta|\theta_1)$ and $f_2^{\eta}(\eta|\theta_2)$ distributions of $\eta(\mathbf{z})$

Limiting behaviour of B_{12} ($\epsilon \to 0$)

When ϵ goes to zero,

$$B^{\eta}_{12}(\mathbf{y}) = \frac{\int \pi_1(\theta_1) f_1^{\eta}(\eta(\mathbf{y})|\theta_1)\,\mathrm{d}\theta_1}{\int \pi_2(\theta_2) f_2^{\eta}(\eta(\mathbf{y})|\theta_2)\,\mathrm{d}\theta_2}$$

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

Limiting behaviour of B_{12} ($\epsilon \rightarrow 0$)

When ϵ goes to zero,

$$B^{\eta}_{12}(\mathbf{y}) = \frac{\int \pi_1(\theta_1) f_1^{\eta}(\eta(\mathbf{y})|\theta_1)\,\mathrm{d}\theta_1}{\int \pi_2(\theta_2) f_2^{\eta}(\eta(\mathbf{y})|\theta_2)\,\mathrm{d}\theta_2}$$

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

Limiting behaviour of B_{12} (under sufficiency)

If $\eta(\mathbf{y})$ sufficient statistic in both models,

$$f_i(\mathbf{y}|\theta_i) = g_i(\mathbf{y}) f_i^\eta(\eta(\mathbf{y})|\theta_i)$$

Thus

$$\begin{aligned} B_{12}(\mathbf{y}) &= \frac{\int_{\Theta_1} \pi(\theta_1) g_1(\mathbf{y}) f_1^\eta(\eta(\mathbf{y})|\theta_1)\,\mathrm{d}\theta_1}{\int_{\Theta_2} \pi(\theta_2) g_2(\mathbf{y}) f_2^\eta(\eta(\mathbf{y})|\theta_2)\,\mathrm{d}\theta_2} \\ &= \frac{g_1(\mathbf{y}) \int \pi_1(\theta_1) f_1^\eta(\eta(\mathbf{y})|\theta_1)\,\mathrm{d}\theta_1}{g_2(\mathbf{y}) \int \pi_2(\theta_2) f_2^\eta(\eta(\mathbf{y})|\theta_2)\,\mathrm{d}\theta_2} = \frac{g_1(\mathbf{y})}{g_2(\mathbf{y})} B_{12}^\eta(\mathbf{y})\,. \end{aligned}$$

[Didelot, Everitt, Johansen & Lawson, 2011]

© No discrepancy only when cross-model sufficiency

Limiting behaviour of B_{12} (under sufficiency)

If $\eta(\mathbf{y})$ sufficient statistic in both models,

$$f_i(\mathbf{y}|\theta_i) = g_i(\mathbf{y}) f_i^\eta(\eta(\mathbf{y})|\theta_i)$$

Thus

$$\begin{aligned} B_{12}(\mathbf{y}) &= \frac{\int_{\Theta_1} \pi(\theta_1) g_1(\mathbf{y}) f_1^\eta(\eta(\mathbf{y})|\theta_1)\,\mathrm{d}\theta_1}{\int_{\Theta_2} \pi(\theta_2) g_2(\mathbf{y}) f_2^\eta(\eta(\mathbf{y})|\theta_2)\,\mathrm{d}\theta_2} \\ &= \frac{g_1(\mathbf{y})\ \int \pi_1(\theta_1) f_1^\eta(\eta(\mathbf{y})|\theta_1)\,\mathrm{d}\theta_1}{g_2(\mathbf{y})\ \int \pi_2(\theta_2) f_2^\eta(\eta(\mathbf{y})|\theta_2)\,\mathrm{d}\theta_2} = \frac{g_1(\mathbf{y})}{g_2(\mathbf{y})} B_{12}^\eta(\mathbf{y})\,. \end{aligned}$$

[Didelot, Everitt, Johansen & Lawson, 2011]

© No discrepancy only when cross-model sufficiency

Poisson/geometric example (back)

Sample

$$\boldsymbol{x} = (x_1, \ldots, x_n)$$

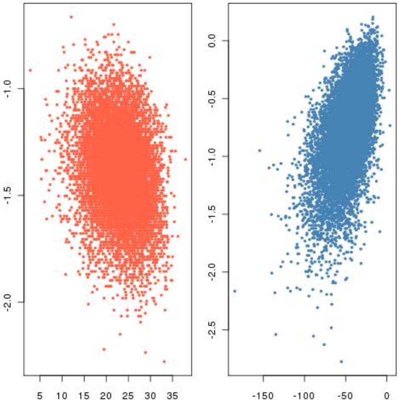
from either a Poisson $\mathcal{P}(\lambda)$ or from a geometric $\mathcal{G}(p)$

Discrepancy ratio

$$\frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})} = \frac{S!n^{-S} / \prod_i x_i!}{1 / \binom{n+S-1}{S}}$$

Poisson/geometric discrepancy

Range of $B_{12}(\boldsymbol{x})$ versus $B^{\eta}_{12}(\boldsymbol{x})$: The values produced have nothing in common.



Formal recovery

Creating an encompassing exponential family

$$f(\boldsymbol{x}|\theta_1, \theta_2, \alpha_1, \alpha_2) \propto \exp\{\theta_1^{\mathsf{T}}\eta_1(\boldsymbol{x}) + \theta_2^{\mathsf{T}}\eta_2(\boldsymbol{x}) + \alpha_1 t_1(\boldsymbol{x}) + \alpha_2 t_2(\boldsymbol{x})\}$$

leads to a sufficient statistic $(\eta_1(\boldsymbol{x}), \eta_2(\boldsymbol{x}), t_1(\boldsymbol{x}), t_2(\boldsymbol{x}))$

[Didelot, Everitt, Johansen & Lawson, 2011]

Formal recovery

Creating an encompassing exponential family

$$f(\mathbf{x}|\theta_1, \theta_2, \alpha_1, \alpha_2) \propto \exp\{\theta_1^{\mathsf{T}}\eta_1(\mathbf{x}) + \theta_2^{\mathsf{T}}\eta_2(\mathbf{x}) + \alpha_1 t_1(\mathbf{x}) + \alpha_2 t_2(\mathbf{x})\}$$

leads to a sufficient statistic $(\eta_1(\mathbf{x}), \eta_2(\mathbf{x}), t_1(\mathbf{x}), t_2(\mathbf{x}))$
[Didelot, Everitt, Johansen & Lawson, 2011]

In the Poisson/geometric case, if $\prod_i x_i!$ is added to S, no discrepancy

Formal recovery

Creating an encompassing exponential family

$$f(\mathbf{x}|\theta_1, \theta_2, \alpha_1, \alpha_2) \propto \exp\{\theta_1^\mathsf{T}\eta_1(\mathbf{x}) + \theta_2^\mathsf{T}\eta_2(\mathbf{x}) + \alpha_1 t_1(\mathbf{x}) + \alpha_2 t_2(\mathbf{x})\}$$

leads to a sufficient statistic $(\eta_1(\mathbf{x}), \eta_2(\mathbf{x}), t_1(\mathbf{x}), t_2(\mathbf{x}))$
[Didelot, Everitt, Johansen & Lawson, 2011]

Only applies in genuine sufficiency settings...

© Inability to evaluate information loss due to summary statistics

Meaning of the ABC-Bayes factor

The ABC approximation to the Bayes Factor is based solely on the summary statistics....

In the Poisson/geometric case, if $\mathbb{E}[y_i] = \theta_0 > 0$,

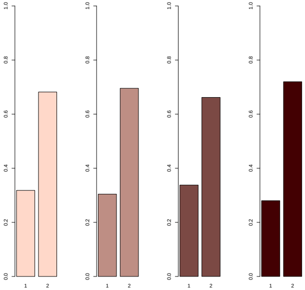
$$\lim_{n\to\infty} B_{12}^{\eta}(\mathbf{y}) = \frac{(\theta_0+1)^2}{\theta_0} e^{-\theta_0}$$

Meaning of the ABC-Bayes factor

The ABC approximation to the Bayes Factor is based solely on the summary statistics....
In the Poisson/geometric case, if $\mathbb{E}[y_i] = \theta_0 > 0$,

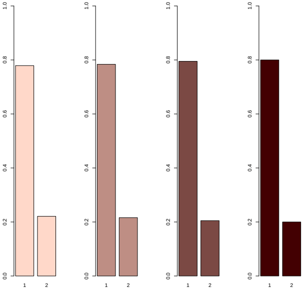
$$\lim_{n \to \infty} B_{12}^{\eta}(\mathbf{y}) = \frac{(\theta_0 + 1)^2}{\theta_0} e^{-\theta_0}$$

MA example



Evolution [against ϵ] of ABC Bayes factor, in terms of frequencies of visits to models MA(1) (left) and MA(2) (right) when ϵ equal to $10, 1, .1, .01\%$ quantiles on insufficient autocovariance distances. Sample of 50 points from a MA(2) with $\theta_1 = 0.6$, $\theta_2 = 0.2$. True Bayes factor equal to 17.71.

MA example



Evolution [against ϵ] of ABC Bayes factor, in terms of frequencies of visits to models MA(1) (left) and MA(2) (right) when ϵ equal to $10, 1, .1, .01\%$ quantiles on insufficient autocovariance distances. Sample of 50 points from a MA(1) model with $\theta_1 = 0.6$. True Bayes factor B_{21} equal to .004.

The only safe cases??? [circa April 2011]

Besides specific models like Gibbs random fields,

using distances over the data itself escapes the discrepancy...
[Toni & Stumpf, 2010; Sousa & al., 2009]

...and so does the use of more informal model fitting measures
[Ratmann & al., 2009]

...or use another type of approximation like empirical likelihood
[Mengersen et al., 2012, see Kerrie's ASC 2012 talk]

The only safe cases??? [circa April 2011]

Besides specific models like Gibbs random fields,

using distances over the data itself escapes the discrepancy...
[Toni & Stumpf, 2010; Sousa & al., 2009]

...and so does the use of more informal model fitting measures
[Ratmann & al., 2009]

...or use another type of approximation like empirical likelihood
[Mengersen et al., 2012, see Kerrie's ASC 2012 talk]

The only safe cases??? [circa April 2011]

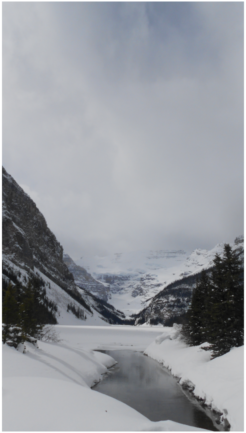
Besides specific models like Gibbs random fields,

using distances over the data itself escapes the discrepancy...
[Toni & Stumpf, 2010; Sousa & al., 2009]

...and so does the use of more informal model fitting measures
[Ratmann & al., 2009]

...or use another type of approximation like empirical likelihood
[Mengersen et al., 2012, see Kerrie's ASC 2012 talk]

ABC model choice consistency



The starting point

Central question to the validation of ABC for model choice:

When is a Bayes factor based on an insufficient statistic $\mathbf{T}(\mathfrak{y})$ consistent?

Note/warnin: © drawn on $\mathbf{T}(\mathfrak{y})$ through $B^{\mathbf{T}}_{12}(\mathfrak{y})$ necessarily differs from © drawn on $\mathfrak{y}$ through $B_{12}(\mathfrak{y})$

[Marin, Pillai, X, & Rousseau, arXiv, 2012]

The starting point

Central question to the validation of ABC for model choice:

When is a Bayes factor based on an insufficient statistic $\mathbf{T}(\mathfrak{y})$ consistent?

Note/warnin: ⓒ drawn on $\mathbf{T}(\mathfrak{y})$ through $B_{12}^{\mathbf{T}}(\mathfrak{y})$ necessarily differs from ⓒ drawn on $\mathfrak{y}$ through $B_{12}(\mathfrak{y})$

[Marin, Pillai, X, & Rousseau, arXiv, 2012]

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks!]:

[X, Cornuet, Marin, & Pillai, Aug. 2011]

Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks!]:

[X, Cornuet, Marin, & Pillai, Aug. 2011]

Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).

Four possible statistics

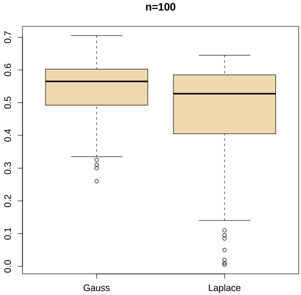
1. sample mean $\overline{\mathbf{y}}$ (sufficient for $\mathfrak{M}_1$ if not $\mathfrak{M}_2$);
2. sample median $\text{med}(\mathbf{y})$ (insufficient);
3. sample variance $\text{var}(\mathbf{y})$ (ancillary);
4. median absolute deviation $\text{mad}(\mathbf{y}) = \text{med}(|\mathbf{y} - \text{med}(\mathbf{y})|)$;

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks!]:

[X, Cornuet, Marin, & Pillai, Aug. 2011]

Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).

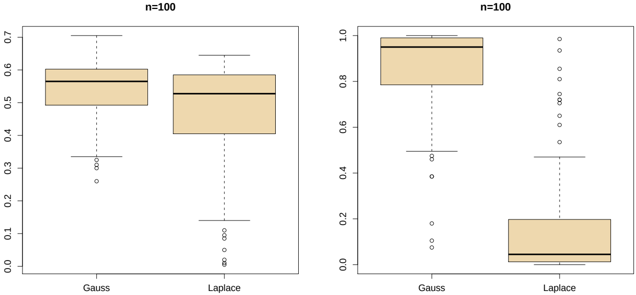


A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks!]:

[X, Cornuet, Marin, & Pillai, Aug. 2011]

Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).



Framework

Starting from sample

$$\mathfrak{y} = (y_1, \ldots, y_n)$$

the observed sample, not necessarily iid with *true* distribution

$$\mathfrak{y} \sim \mathbb{P}^n$$

Summary statistics

$$\mathbf{T}(\mathfrak{y}) = \mathbf{T}^n = (T_1(\mathfrak{y}), T_2(\mathfrak{y}), \cdots, T_d(\mathfrak{y})) \in \mathbb{R}^d$$

with *true* distribution $\mathbf{T}^n \sim G_n$.

Framework

© Comparison of

- under $\mathfrak{M}_1$, $\mathfrak{y} \sim F_{1,n}(\cdot|\theta_1)$ where $\theta_1 \in \Theta_1 \subset \mathbb{R}^{p_1}$
- under $\mathfrak{M}_2$, $\mathfrak{y} \sim F_{2,n}(\cdot|\theta_2)$ where $\theta_2 \in \Theta_2 \subset \mathbb{R}^{p_2}$

turned into

- under $\mathfrak{M}_1$, $\mathbf{T}(\mathfrak{y}) \sim G_{1,n}(\cdot|\theta_1)$, and $\theta_1|\mathbf{T}(\mathfrak{y}) \sim \pi_1(\cdot|\mathbf{T}^n)$
- under $\mathfrak{M}_2$, $\mathbf{T}(\mathfrak{y}) \sim G_{2,n}(\cdot|\theta_2)$, and $\theta_2|\mathbf{T}(\mathfrak{y}) \sim \pi_2(\cdot|\mathbf{T}^n)$

Assumptions

A collection of asymptotic "standard" assumptions:

[A1] is a standard central limit theorem under the true model
[A2] controls the large deviations of the estimator $\mathbf{T}^n$ from the estimand $\mu(\theta)$
[A3] is the standard prior mass condition found in Bayesian asymptotics (d_i effective dimension of the parameter)
[A4] restricts the behaviour of the model density against the true density

[Think CLT!]

Assumptions

A collection of asymptotic "standard" assumptions:

[Think CLT!]

[A1] There exist

- a sequence $\{v_n\}$ converging to $+\infty$,
- a distribution Q,
- a symmetric, $d \times d$ positive definite matrix V_0 and
- a vector $\mu_0 \in \mathbb{R}^d$

such that

$$v_n V_0^{-1/2}(\mathbf{T}^n - \mu_0) \overset{n \to \infty}{\rightsquigarrow} Q, \text{ under } G_n$$

Assumptions

A collection of asymptotic “standard” assumptions:

[Think CLT!]

[A2] For $i = 1, 2$, there exist sets $\mathcal{F}_{n,i} \subset \Theta_i$, functions $\mu_i(\theta_i)$ and constants $\epsilon_i, \tau_i, \alpha_i > 0$ such that for all $\tau > 0$,

$$\sup_{\theta_i \in \mathcal{F}_{n,i}} \frac{G_{i,n}\left[|\mathbf{T}^n - \mu_i(\theta_i)| > \tau|\mu_i(\theta_i) - \mu_0| \wedge \epsilon_i \,|\theta_i\right]}{(|\tau\mu_i(\theta_i) - \mu_0| \wedge \epsilon_i)^{-\alpha_i}} \lesssim v_n^{-\alpha_i}$$

with

$$\pi_i(\mathcal{F}_{n,i}^c) = o(v_n^{-\tau_i}).$$

Assumptions

A collection of asymptotic "standard" assumptions:

[Think CLT!]

[A3] If $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, defining $(u > 0)$

$$S_{n,i}(u) = \left\{\theta_i \in \mathcal{F}_{n,i}; |\mu_i(\theta_i) - \mu_0| \leqslant u\, v_n^{-1}\right\},$$

there exist constants $d_i < \tau_i \wedge \alpha_i - 1$ such that

$$\pi_i(S_{n,i}(u)) \sim u^{d_i} v_n^{-d_i}, \quad \forall u \lesssim v_n$$

Assumptions

A collection of asymptotic "standard" assumptions:

[Think CLT!]

[A4] If $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, for any $\epsilon > 0$, there exist $U, \delta > 0$ and $(E_n)_n$ such that, if $\theta_i \in S_{n,i}(U)$

$$E_n \subset \{t; g_i(t|\theta_i) < \delta g_n(t)\} \quad \text{and} \quad G_n\left(E_n^c\right) < \epsilon.$$

Assumptions

A collection of asymptotic “standard” assumptions:

[Think CLT!]

Again (sumup)
[A1] is a standard central limit theorem under the true model
[A2] controls the large deviations of the estimator $\mathbf{T}^n$ from the estimand $\mu(\theta)$
[A3] is the standard prior mass condition found in Bayesian asymptotics (d_i effective dimension of the parameter)
[A4] restricts the behaviour of the model density against the true density

Effective dimension

Understanding d_i in **[A3]**:
defined only when $\mu_0 \in \{\mu_i(\theta_i), \theta_i \in \Theta_i\}$,

$$\pi_i(\theta_i : |\mu_i(\theta_i) - \mu_0| < n^{-1/2}) = O(n^{-d_i/2})$$

is the effective dimension of the model Θ_i around μ_0

Effective dimension

Understanding d_i in **[A3]**:
when $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$,

$$\frac{m_i(\mathbf{T}^n)}{g_n(\mathbf{T}^n)} \sim v_n^{-d_i} \,,$$

thus

$$\log(m_i(\mathbf{T}^n)/g_n(\mathbf{T}^n)) \sim -d_i \log v_n$$

and $v_n^{-d_i}$ penalization factor resulting from integrating θ_i out (see effective number of parameters in DIC)

Effective dimension

Understanding d_i in **[A3]**:
In regular models, d_i dimension of $\mathbf{T}(\Theta_i)$, leading to BIC
In non-regular models, d_i can be smaller

Asymptotic marginals

Asymptotically, under **[A1]–[A4]**

$$m_i(t) = \int_{\Theta_i} g_i(t|\theta_i)\,\pi_i(\theta_i)\,\mathrm{d}\theta_i$$

is such that
(i) if $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$,

$$C_l v_n^{d-d_i} \leqslant m_i(\mathbf{T}^n) \leqslant C_u v_n^{d-d_i}$$

and
(ii) if $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} > 0$

$$m_i(\mathbf{T}^n) = o_{\mathbb{P}^n}[v_n^{d-\tau_i} + v_n^{d-\alpha_i}].$$

Within-model consistency

Under same assumptions,

$$\text{if } \inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0,$$

the posterior distribution of $\mu_i(\theta_i)$ given $\mathbf{T}^n$ is consistent at rate $1/v_n$ provided $\alpha_i \wedge \tau_i > d_i$.

Between-model consistency

Consequence of above is that asymptotic behaviour of the Bayes factor is driven by the asymptotic mean value of $\mathbf{T}^n$ under both models. **And only by this mean value!**

Between-model consistency

Consequence of above is that asymptotic behaviour of the Bayes factor is driven by the asymptotic mean value of $\mathbf{T}^n$ under both models. **And only by this mean value!**

Indeed, if

$$\inf\{|\mu_0 - \mu_2(\theta_2)|; \theta_2 \in \Theta_2\} = \inf\{|\mu_0 - \mu_1(\theta_1)|; \theta_1 \in \Theta_1\} = 0$$

then

$$C_l v_n^{-(d_1 - d_2)} \leqslant m_1(\mathbf{T}^n) \big/ m_2(\mathbf{T}^n) \leqslant C_u v_n^{-(d_1 - d_2)},$$

where $C_l, C_u = O_{\mathbb{P}^n}(1)$, irrespective of the true model.
Ⓒ Only depends on the difference $d_1 - d_2$: no consistency

Between-model consistency

Consequence of above is that asymptotic behaviour of the Bayes factor is driven by the asymptotic mean value of $\mathbf{T}^n$ under both models. **And only by this mean value!**

Else, if

$$\inf\{|\mu_0 - \mu_2(\theta_2)|; \theta_2 \in \Theta_2\} > \inf\{|\mu_0 - \mu_1(\theta_1)|; \theta_1 \in \Theta_1\} = 0$$

then

$$\frac{m_1(\mathbf{T}^n)}{m_2(\mathbf{T}^n)} \geqslant C_u \min\left(v_n^{-(d_1-\alpha_2)}, v_n^{-(d_1-\tau_2)} \right)$$

Consistency theorem

If

$$\inf\{|\mu_0 - \mu_2(\theta_2)|; \theta_2 \in \Theta_2\} = \inf\{|\mu_0 - \mu_1(\theta_1)|; \theta_1 \in \Theta_1\} = 0,$$

Bayes factor

$$B_{12}^{\mathbf{T}} = \mathrm{O}(v_n^{-(d_1 - d_2)})$$

irrespective of the true model. It is inconsistent since it always picks the model with the smallest dimension

Consistency theorem

If $\mathbb{P}^n$ belongs to one of the two models and if μ_0 cannot be attained by the other one :

$$\begin{aligned} 0 &= \min\left(\inf\{|\mu_0 - \mu_i(\theta_i)|; \theta_i \in \Theta_i\}, i = 1, 2\right) \\ &< \max\left(\inf\{|\mu_0 - \mu_i(\theta_i)|; \theta_i \in \Theta_i\}, i = 1, 2\right), \end{aligned}$$

then the Bayes factor $B_{12}^{\mathbf{T}}$ is consistent

Consequences on summary statistics

Bayes factor driven by the means $\mu_i(\theta_i)$ and the relative position of μ_0 wrt both sets $\{\mu_i(\theta_i); \theta_i \in \Theta_i\}$, $i = 1, 2$.

For ABC, this implies the most likely statistics $\mathbf{T}^n$ are ancillary statistics with different mean values under both models

Else, if $\mathbf{T}^n$ asymptotically depends on some of the parameters of the models, it is possible that there exists $\theta_i \in \Theta_i$ such that $\mu_i(\theta_i) = \mu_0$ even though model $\mathfrak{M}_1$ is misspecified

Consequences on summary statistics

Bayes factor driven by the means $\mu_i(\theta_i)$ and the relative position of μ_0 wrt both sets $\{\mu_i(\theta_i); \theta_i \in \Theta_i\}$, $i = 1, 2$.

For ABC, this implies the most likely statistics $\mathbf{T}^n$ are ancillary statistics with different mean values under both models

Else, if $\mathbf{T}^n$ asymptotically depends on some of the parameters of the models, it is possible that there exists $\theta_i \in \Theta_i$ such that $\mu_i(\theta_i) = \mu_0$ even though model $\mathfrak{M}_1$ is misspecified

Toy example: Laplace versus Gauss [1]

If

$$\mathbf{T}^n = n^{-1}\sum_{i=1}^{n} X_i^4, \quad \mu_1(\theta) = 3 + \theta^4 + 6\theta^2, \quad \mu_2(\theta) = 6 + \cdots$$

and the true distribution is Laplace with mean $\theta_0 = 1$, under the Gaussian model the value $\theta^* = 2\sqrt{3} - 3$ also leads to $\mu_0 = \mu(\theta^*)$

[here $d_1 = d_2 = d = 1$]

Toy example: Laplace versus Gauss [1]

If

$$\mathbf{T}^n = n^{-1}\sum_{i=1}^{n} X_i^4, \quad \mu_1(\theta) = 3 + \theta^4 + 6\theta^2, \quad \mu_2(\theta) = 6 + \cdots$$

and the true distribution is Laplace with mean $\theta_0 = 1$, under the Gaussian model the value $\theta^* = 2\sqrt{3} - 3$ also leads to $\mu_0 = \mu(\theta^*)$

[here $d_1 = d_2 = d = 1$]

© A Bayes factor associated with such a statistic is inconsistent

Toy example: Laplace versus Gauss [1]

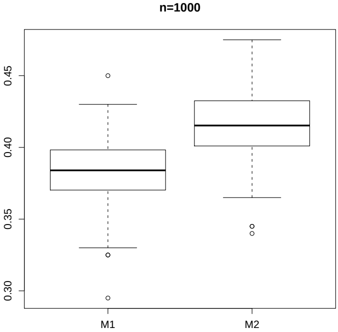
If

$$\mathbf{T}^n = n^{-1} \sum_{i=1}^{n} X_i^4, \quad \mu_1(\theta) = 3 + \theta^4 + 6\theta^2, \quad \mu_2(\theta) = 6 + \cdots$$

Toy example: Laplace versus Gauss [1]

If

$$\mathbf{T}^n = n^{-1}\sum_{i=1}^{n} X_i^4, \quad \mu_1(\theta) = 3 + \theta^4 + 6\theta^2, \quad \mu_2(\theta) = 6 + \cdots$$



Fourth moment

Toy example: Laplace versus Gauss [0]

When

$$\mathbf{T}(\mathfrak{y}) = \left\{ \bar{y}_n^{(4)}, \bar{y}_n^{(6)} \right\}$$

and the true distribution is Laplace with mean $\theta_0 = 0$, then $\mu_0 = 6$, $\mu_1(\theta_1^*) = 6$ with $\theta_1^* = 2\sqrt{3} - 3$

$[d_1 = 1$ and $d_2 = 1/2]$

thus

$$B_{12} \sim n^{-1/4} \rightarrow 0 : \quad \text{consistent}$$

Under the Gaussian model $\mu_0 = 3$, $\mu_2(\theta_2) \geqslant 6 > 3\ \forall \theta_2$

$$B_{12} \rightarrow +\infty : \quad \text{consistent}$$

Toy example: Laplace versus Gauss [0]

When

$$\mathbf{T}(\mathfrak{y}) = \left\{ \bar{y}_n^{(4)}, \bar{y}_n^{(6)} \right\}$$

and the true distribution is Laplace with mean $\theta_0 = 0$, then $\mu_0 = 6$, $\mu_1(\theta_1^*) = 6$ with $\theta_1^* = 2\sqrt{3} - 3$

$[d_1 = 1 \text{ and } d_2 = 1/2]$

thus

$$B_{12} \sim n^{-1/4} \to 0 : \quad \text{consistent}$$

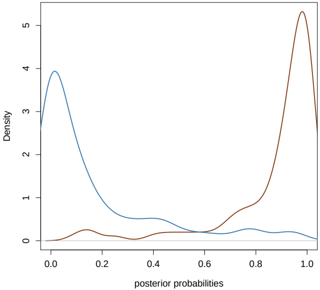
Under the Gaussian model $\mu_0 = 3$, $\mu_2(\theta_2) \geqslant 6 > 3 \ \forall \theta_2$

$$B_{12} \to +\infty : \quad \text{consistent}$$

Toy example: Laplace versus Gauss [0]

When

$$\mathbf{T}(\mathfrak{y}) = \left\{ \bar{y}_n^{(4)}, \bar{y}_n^{(6)} \right\}$$



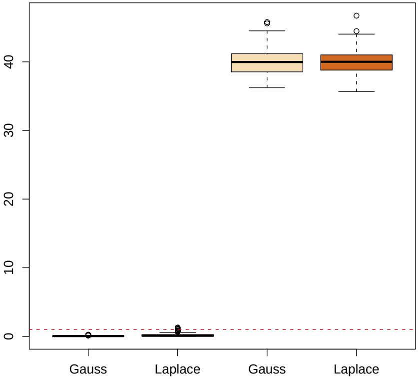
Fourth AND **sixth moments**

Checking for adequate statistics

After running ABC, i.e. creating *reference tables* of (θ_i, x_i) from both joints, straightforward derivation of ABC estimates $\hat{\theta}_1$ and $\hat{\theta}_2$.

Evaluation of $\mathbb{E}^1_{\hat{\theta}_1}[T(X)]$ and $\mathbb{E}^2_{\hat{\theta}_2}[T(X)]$ allows for detection of different means under both models via Monte Carlo simulations

Toy example: Laplace versus Gauss

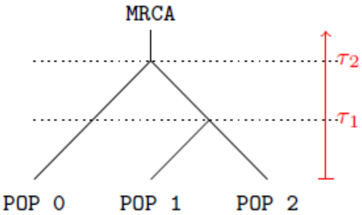


Normalised χ^2 without and with mad

A population genetic illustration

Two populations (1 and 2) having diverged at a fixed known time in the past and third population (3) which diverged from one of those two populations (models 1 and 2, respectively).

Observation of 50 diploid individuals/population genotyped at 5, 50 or 100 independent microsatellite loci.



Model 2

A population genetic illustration

Two populations (1 and 2) having diverged at a fixed known time in the past and third population (3) which diverged from one of those two populations (models 1 and 2, respectively).

Observation of 50 diploid individuals/population genotyped at 5, 50 or 100 independent microsatellite loci.

Stepwise mutation model: the number of repeats of the mutated gene increases or decreases by one. Mutation rate μ common to all loci set to 0.005 (single parameter) with uniform prior distribution

$$\mu \sim \mathcal{U}[0.0001, 0.01]$$

A population genetic illustration

Summary statistics associated to the $(\delta_\mu)^2$ distance

$x_{l,i,j}$ repeated number of allele in locus $l = 1, \ldots, L$ for individual $i = 1, \ldots, 100$ within the population $j = 1, 2, 3$. Then

$$(\delta_\mu)^2_{j_1,j_2} = \frac{1}{L} \sum_{l=1}^{L} \left(\frac{1}{100} \sum_{i_1=1}^{100} x_{l,i_1,j_1} - \frac{1}{100} \sum_{i_2=1}^{100} x_{l,i_2,j_2} \right)^2 .$$

A population genetic illustration

For two copies of locus l with allele sizes x_{l,i,j_1} and x_{l,i',j_2}, most recent common ancestor at coalescence time τ_{j_1,j_2}, gene genealogy distance of $2\tau_{j_1,j_2}$, hence number of mutations Poisson with parameter $2\mu\tau_{j_1,j_2}$. Therefore,

$$\mathbb{E}\left\{\left(x_{l,i,j_1} - x_{l,i',j_2}\right)^2 |\tau_{j_1,j_2}\right\} = 2\mu\tau_{j_1,j_2}$$

and

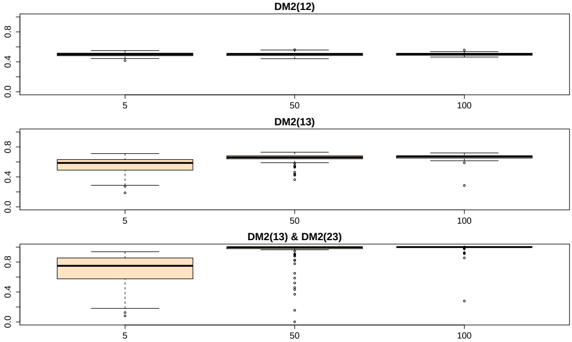
	Model 1	Model 2
$\mathbb{E}\left\{(\delta_\mu)^2_{1,2}\right\}$	$2\mu_1 t'$	$2\mu_2 t'$
$\mathbb{E}\left\{(\delta_\mu)^2_{1,3}\right\}$	$2\mu_1 t$	$2\mu_2 t'$
$\mathbb{E}\left\{(\delta_\mu)^2_{2,3}\right\}$	$2\mu_1 t'$	$2\mu_2 t$

A population genetic illustration

Thus,

- Bayes factor based only on distance $(\delta_\mu)^2_{1,2}$ not convergent: if $\mu_1 = \mu_2$, same expectation
- Bayes factor based only on distance $(\delta_\mu)^2_{1,3}$ or $(\delta_\mu)^2_{2,3}$ not convergent: if $\mu_1 = 2\mu_2$ or $2\mu_1 = \mu_2$ same expectation
- if two of the three distances are used, Bayes factor converges: there is no (μ_1, μ_2) for which all expectations are equal

A population genetic illustration



Posterior probabilities that the data is from model 1 for 5, 50 and 100 loci

Embedded models

When $\mathfrak{M}_1$ submodel of $\mathfrak{M}_2$, and if the true distribution belongs to the smaller model $\mathfrak{M}_1$, Bayes factor is of order

$$v_n^{-(d_1-d_2)}$$

Embedded models

When $\mathfrak{M}_1$ submodel of $\mathfrak{M}_2$, and if the true distribution belongs to the smaller model $\mathfrak{M}_1$, Bayes factor is of order

$$v_n^{-(d_1-d_2)}$$

If summary statistic only informative on a parameter that is the same under both models, i.e if $d_1 = d_2$, then

ⓒ the Bayes factor is not consistent

Embedded models

When $\mathfrak{M}_1$ submodel of $\mathfrak{M}_2$, and if the true distribution belongs to the smaller model $\mathfrak{M}_1$, Bayes factor is of order

$$v_n^{-(d_1-d_2)}$$

Else, $d_1 < d_2$ and Bayes factor is going to ∞ under $\mathfrak{M}_1$. If true distribution not in $\mathfrak{M}_1$, then

ⓒ Bayes factor is consistent only if $\mu_1 \neq \mu_2 = \mu_0$

Summary

- Model selection feasible with ABC
- Choice of summary statistics is paramount
- At best, ABC $\to \pi(. \mid \mathbf{T}(y))$ which concentrates around μ_0
- For estimation : $\{\theta; \mu(\theta) = \mu_0\} = \theta_0$
- For testing $\{\mu_1(\theta_1), \theta_1 \in \Theta_1\} \cap \{\mu_2(\theta_2), \theta_2 \in \Theta_2\} = \emptyset$

Summary

- Model selection feasible with ABC
- Choice of summary statistics is paramount
- At best, ABC $\to \pi(. \mid \mathbf{T}(y))$ which concentrates around μ_0
- For estimation : $\{\theta; \mu(\theta) = \mu_0\} = \theta_0$
- For testing $\{\mu_1(\theta_1), \theta_1 \in \Theta_1\} \cap \{\mu_2(\theta_2), \theta_2 \in \Theta_2\} = \emptyset$

Empirical likelihood (EL)

Empirical likelihood (EL)

▸ help!

Dataset x made of n independent replicates $x = (x_1, \dots, x_n)$ of some $X \sim F$

Generalized moment condition model

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

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

Corresponding empirical likelihood

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

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

[Owen, 1988, Bio'ka; Owen, 2001]

Empirical likelihood (EL)

help!

Dataset x made of n independent replicates $x = (x_1, \ldots, x_n)$ of some $X \sim F$

Generalized moment condition model

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

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

Corresponding empirical likelihood

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

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

[Owen, 1988, Bio'ka; Owen, 2001]

Convergence of EL [3.4]

Theorem 3.4 *Let* $X, Y_1, \ldots, Y_n$ *be independent rv's with common distribution* F_0. *For* $\theta \in \Theta$, *and the function* $h(X, \theta) \in \mathbb{R}^s$, *let* $\theta_0 \in \Theta$ *be such that*

$$\mathrm{Var}(h(Y_i, \theta_0))$$

is finite and has rank $q > 0$. *If* θ_0 *satisfies*

$$\mathbb{E}(h(X, \theta_0)) = 0,$$

then

$$-2 \log \left(\frac{L_{\mathrm{el}}(\theta_0 | Y_1, \ldots, Y_n)}{n^{-n}} \right) \to \chi^2_{(q)}$$

in distribution when $n \to \infty$.

[Owen, 2001]

Convergence of EL [3.4]

"...The interesting thing about Theorem 3.4 is what is not there. It includes no conditions to make $\widehat{\theta}$ a good estimate of θ_0, nor even conditions to ensure a unique value for θ_0, nor even that any solution θ_0 exists. Theorem 3.4 applies in the just determined, over-determined, and under-determined cases. When we can prove that our estimating equations uniquely define θ_0, and provide a consistent estimator $\widehat{\theta}$ of it, then confidence regions and tests follow almost automatically through Theorem 3.4.".

[Owen, 2001]

Raw ABC$_{\text{el}}$sampler

Act as if EL was an exact likelihood

[Lazar, 2003]

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

- Output weighted sample of size N

Raw ABC_{el}sampler

Act as if EL was an exact likelihood

[Lazar, 2003]

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

- Performance evaluated through effective sample size

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

Raw ABC$_{el}$ sampler

Act as if EL was an exact likelihood

[Lazar, 2003]

```
for $i = 1 \rightarrow N$ do
    generate $\phi_i$ from the prior distribution $\pi(\cdot)$
    set the weight $\omega_i = L_{el}(\phi_i|x_{obs})$
end for
return $(\phi_i, \omega_i)$, $i = 1, \dots, N$
```

- More advanced algorithms can be adapted to EL: E.g., adaptive multiple importance sampling (AMIS) of Cornuet et al. to speed up computations

[Cornuet et al., 2012]

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 made of the genotypes of the sample of individuals at a given locus

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

- dates of divergence between 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 made of the genotypes of the sample of individuals at a given locus

© h made of 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 $x_k^1, \dots, x_k^n$: 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 narrower than the original likelihood 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 $\delta = x_k^i - x_k^j$

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

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: 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 $\delta = x_k^i - x_k^j$

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

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: 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 $\delta = x_k^i - x_k^j$

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

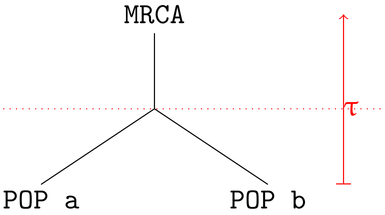
with

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

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

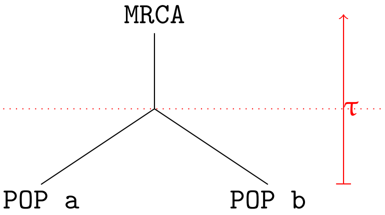
Let x_k^i and x_k^j be two genes coming resp. from pop. *a* and *b*
Set $\delta = x_k^i - x_k^j$.

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

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

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

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

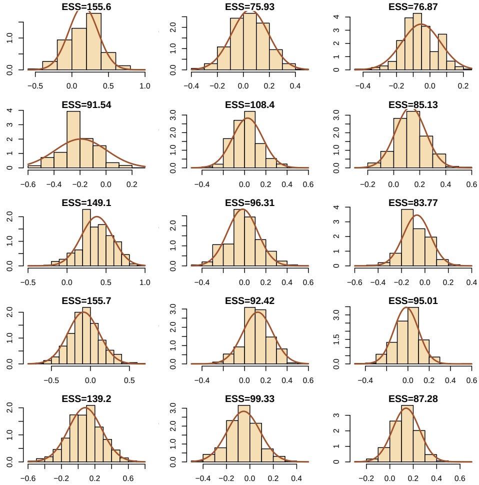
$$\partial_\theta \log \ell_2(\delta|\theta,\tau) = -\tau - \frac{1}{1+2\theta} + \frac{q(\delta|\theta,\tau)}{\ell_2(\delta|\theta,\tau)} + \frac{\tau}{2}\frac{\ell_2(\delta-1|\theta,\tau)+\ell_2(\delta+1|\theta,\tau)}{\ell_2(\delta|\theta,\tau)}$$

where

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

Example: normal posterior

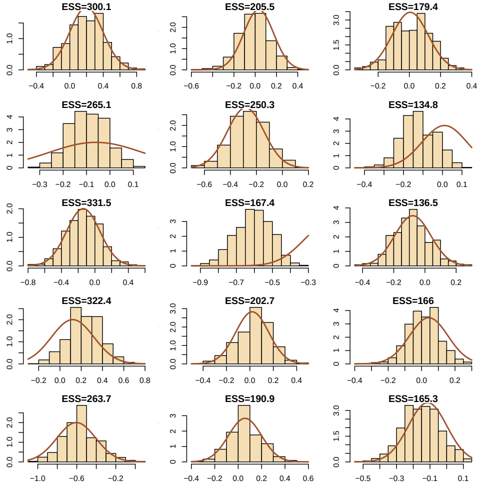
ABC_{el} with two constraints



Sample sizes are of 25 (column 1), 50 (column 2) and 75 (column 3) observations

Example: normal posterior

ABC$_{el}$ with three constraints



Sample sizes are of 25 (column 1), 50 (column 2) and 75 (column 3) observations

Example: Superposition of gamma processes

Example of superposition of N renewal processes with waiting times τ_{ij} $(i = 1, \ldots, M)$, $j = 1, \ldots) \sim \mathcal{G}(\alpha, \beta)$, when N is unknown. Renewal processes

$$\zeta_{i1} = \tau_{i1}, \ \zeta_{i2} = \zeta_{i1} + \tau_{i2}, \ \ldots$$

with observations made of first n values of the ζ_{ij}'s,

$$z_1 = \min\{\zeta_{ij}\}, \ z_2 = \min\{\zeta_{ij}; \zeta_{ij} > z_1\}, \ \ldots,$$

ending with

$$z_n = \min\{\zeta_{ij}; \zeta_{ij} > z_{n-1}\}\,.$$

[Cox & Kartsonaki, B'ka, 2012]

Example: Superposition of gamma processes (ABC)

Interesting testing ground for ABC_{el} since data (z_t) neither iid nor Markov

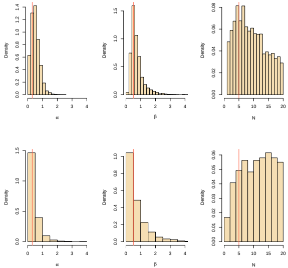
Recovery of an iid structure by

1. simulating a pseudo-dataset, $(z_1^\star, \ldots, z_n^\star)$, as in regular ABC,
2. deriving sequence of indicators $(\nu_1, \ldots, \nu_n)$, as

$$z_1^\star = \zeta_{\nu_1 1},\ z_2^\star = \zeta_{\nu_2 j_2},\ \ldots$$

3. exploiting that those indicators are distributed from the prior distribution on the ν_t's leading to an iid sample of $\mathcal{G}(\alpha, \beta)$ variables

Comparison of ABC and ABC_{el} posteriors



Top: ABC_{el}
Bottom: regular ABC

Example: Superposition of gamma processes (ABC)

Interesting testing ground for ABC$_{\text{el}}$ since data (z_t) neither iid nor Markov

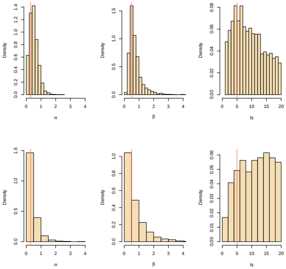
Recovery of an iid structure by

1. simulating a pseudo-dataset, $(z_1^\star, \dots, z_n^\star)$, as in regular ABC,
2. deriving sequence of indicators $(\nu_1, \dots, \nu_n)$, as

$$z_1^\star = \zeta_{\nu_1 1},\ z_2^\star = \zeta_{\nu_2 j_2},\ \dots$$

3. exploiting that those indicators are distributed from the prior distribution on the ν_t's leading to an iid sample of $\mathcal{G}(\alpha, \beta)$ variables

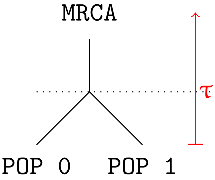
Comparison of ABC and ABC$_{\text{el}}$ posteriors



Top: ABC$_{\text{el}}$
Bottom: regular ABC

Pop'gen': A first experiment

Evolutionary scenario:



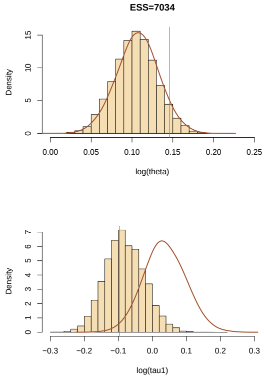
Dataset:

- 50 genes per populations,
- 100 microsat. loci

Assumptions:

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

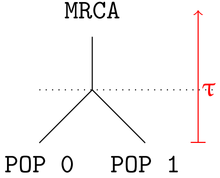
Comparison of the original ABC with ABC$_{\text{el}}$



histogram = ABC$_{\text{el}}$
curve = original ABC
vertical line = "true" parameter

Pop'gen': A first experiment

Evolutionary scenario:



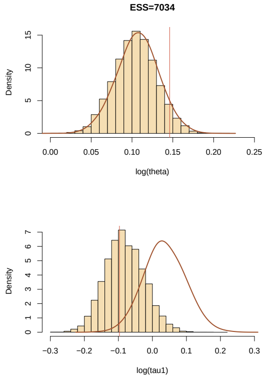
Dataset:

- 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$_{\text{el}}$



histogram = ABC$_{\text{el}}$
curve = original ABC
vertical line = "true" parameter

ABC vs. ABC$_{\text{el}}$ on 100 replicates of the 1st experiment

Accuracy:

	$\log_{10}\theta$		$\log_{10}\tau$	
	ABC	ABC$_{\text{el}}$	ABC	ABC$_{\text{el}}$
(1)	0.097	0.094	0.315	0.117
(2)	0.071	0.059	0.272	0.077
(3)	0.68	0.81	1.0	0.80

(1) Root Mean Square Error of the posterior mean

(2) Median Absolute Deviation of the posterior median

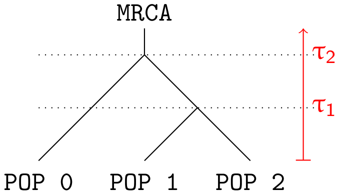
(3) Coverage of the credibility interval of probability 0.8

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

- ABC $\approx$ 4 hours
- ABC$_{\text{el}}\approx$ 2 minutes

Pop'gen': 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 uniform

Comparison of the original ABC with ABC$_{el}$

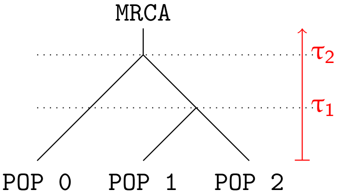
histogram = ABC$_{el}$

curve = original ABC

vertical line = “true” parameter

Pop'gen': Second experiment

Evolutionary scenario:



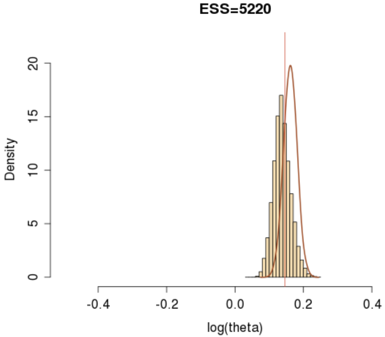
Dataset:

- 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 uniform

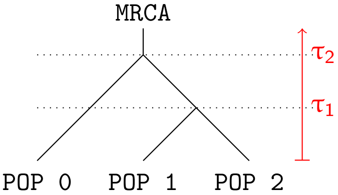
Comparison of the original ABC with ABC$_{\text{el}}$



histogram = ABC$_{\text{el}}$
curve = original ABC
vertical line = "true" parameter

Pop'gen': Second experiment

Evolutionary scenario:



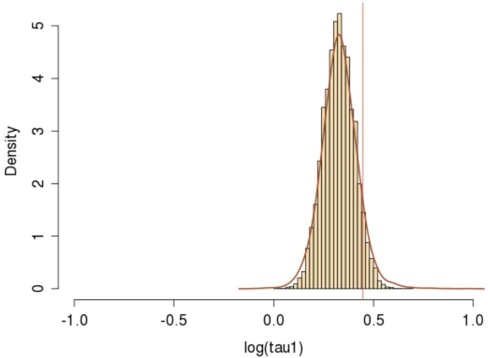
Dataset:

- 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 uniform

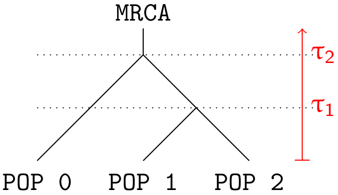
Comparison of the original ABC with ABC$_{\text{el}}$



histogram = ABC$_{\text{el}}$
curve = original ABC
vertical line = "true" parameter

Pop'gen': Second experiment

Evolutionary scenario:



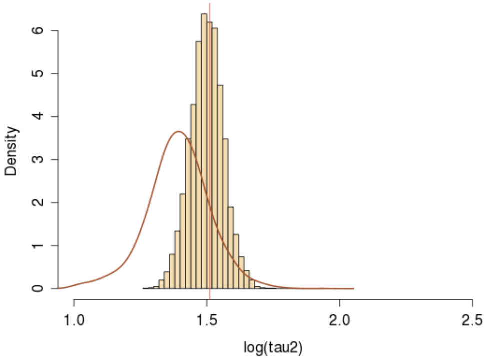
Dataset:

- 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 uniform

Comparison of the original ABC with ABC$_{\text{el}}$



histogram = ABC$_{\text{el}}$
curve = original ABC
vertical line = "true" parameter

ABC vs. ABC$_{\text{el}}$ on 100 replicates of the 2nd experiment

Accuracy:

	$\log_{10}\theta$		$\log_{10}\tau_1$		$\log_{10}\tau_2$	
	ABC	ABC$_{\text{el}}$	ABC	ABC$_{\text{el}}$	ABC	ABC$_{\text{el}}$
(1)	.0059	.0794	.472	.483	29.3	4.76
(2)	.048	.053	.32	.28	4.13	3.36
(3)	.79	.76	.88	.76	.89	.79

(1) Root Mean Square Error of the posterior mean

(2) Median Absolute Deviation of the posterior median

(3) Coverage of the credibility interval of probability 0.8

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

- ABC $\approx$ 6 hours
- ABC$_{\text{el}}\approx$ 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

$$\pi\Big(\theta \,\Big|\, \eta(x_{\text{obs}})\Big),$$

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

ABC_{el}simplifies the model through a generalized moment condition model.
↪ Here, the moment condition model is based on pairwise composition likelihood

Personal Call

My son Joachim, 19, is looking for a summer internship as a salesman in the US in the summer 2013, requirement of his business school (Iéseg) curriculum. Any help in this matter appreciated!