ABC methodology and applications

Christian P. Robert

Université Paris-Dauphine, University of Warwick, & IUF

~~École d'Hiver, Les Diablerets, CH, Feb 4-8 2017~~
~~Astrostat Fall School, Autrans, Oct 11-12 2017~~
~~1st Franco-Italian Workshop, Ca'Foscari, Oct 16-17, 2017~~

Outline

1. simulation-based methods in Econometrics
2. Genetics of ABC
3. Approximate Bayesian computation
4. ABC for model choice
5. ABC model choice via random forests
6. ABC estimation via random forests
7. [some] asymptotics of ABC

A motivating if pedestrian example

paired and orphan socks

A drawer contains an unknown number of socks, some of which can be paired and some of which are orphans (single). One takes at random 11 socks without replacement from this drawer: no pair can be found among those. What can we infer about the total number of socks in the drawer?

A motivating if pedestrian example

paired and orphan socks

A drawer contains an unknown number of socks, some of which can be paired and some of which are orphans (single). One takes at random 11 socks without replacement from this drawer: no pair can be found among those. What can we infer about the total number of socks in the drawer?

- sounds like an impossible task
- one observation $x = 11$ and two unknowns, n_{socks} and n_{pairs}
- writing the likelihood is a challenge [exercise]

Feller's shoes

A closet contains n pairs of shoes. If $2r$ shoes are chosen at random (with $2r < n$), what is the probability that there will be (a) no complete pair, (b) exactly one complete pair, (c) exactly two complete pairs among them?

[Feller, 1970, Chapter II, Exercise 26]

Feller's shoes

> *A closet contains n pairs of shoes. If $2r$ shoes are chosen at random (with $2r < n$), what is the probability that there will be (a) no complete pair, (b) exactly one complete pair, (c) exactly two complete pairs among them?*

[Feller, 1970, Chapter II, Exercise 26]

Resolution as

$$p_j = \binom{n}{j} 2^{2r-2j} \binom{n-j}{2r-2j} \Big/ \binom{2n}{2r}$$

being probability of obtaining js pairs among those $2r$ shoes, or for an odd number t of shoes

$$p_j = 2^{t-2j} \binom{n}{j} \binom{n-j}{t-2j} \Big/ \binom{2n}{t}$$

Feller's shoes

> *A closet contains n pairs of shoes. If $2r$ shoes are chosen at random (with $2r < n$), what is the probability that there will be (a) no complete pair, (b) exactly one complete pair, (c) exactly two complete pairs among them?*

[Feller, 1970, Chapter II, Exercise 26]

If one draws 11 socks out of m socks made of f orphans and g pairs, with $f + 2g = m$, number k of socks from the orphan group is hypergeometric $\mathscr{H}(11, m, f)$ and probability to observe 11 orphan socks total is

$$\sum_{k=0}^{11} \frac{\binom{f}{k}\binom{2g}{11-k}}{\binom{m}{11}} \times \frac{2^{11-k}\binom{g}{11-k}}{\binom{2g}{11-k}}$$

A prioris on socks

Given parameters n_{socks} and n_{pairs}, set of socks

$$\mathcal{S} = \left\{s_1, s_1, \ldots, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}+1}, \ldots, s_{n_{\text{socks}}}\right\}$$

and 11 socks picked at random from $\mathcal{S}$ give X unique socks.

A prioris on socks

Given parameters n_{socks} and n_{pairs}, set of socks

$$\mathcal{S} = \left\{s_1, s_1, \ldots, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}+1}, \ldots, s_{n_{\text{socks}}}\right\}$$

and 11 socks picked at random from $\mathcal{S}$ give X unique socks.

Rassmus' reasoning

If you are a family of 3-4 persons then a guesstimate would be that you have something like 15 pairs of socks in store. It is also possible that you have much more than 30 socks. So as *a prior* for n_{socks} I'm going to use a negative binomial with mean 30 and standard deviation 15.
On ${}^{n_{\text{pairs}}}/_{2n_{\text{socks}}}$ I'm going to put a Beta *prior* distribution that puts most of the probability over the range 0.75 to 1.0,

[Rassmus Bååth's Research Blog, Oct 20th, 2014]

Simulating the experiment

Given a *prior* distribution on n_{socks} and n_{pairs},

$$n_{\text{socks}} \sim \mathcal{N}eg(30,15) \quad n_{\text{pairs}}|n_{\text{socks}} \sim {}^{n_{\text{socks}}}/_{2}\mathcal{B}e(15,2)$$

possible to

1. generate new values of n_{socks} and n_{pairs},
2. generate a new observation of X, number of unique socks out of 11.

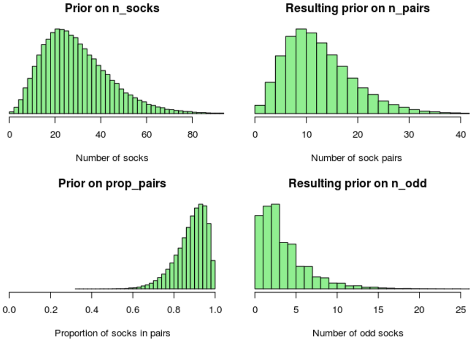
Simulating the experiment

Given a *prior* distribution on n_{socks} and n_{pairs},

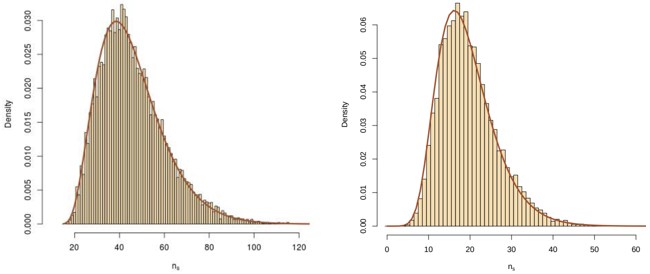
$$n_{\text{socks}} \sim \mathcal{N}eg(30,15) \quad n_{\text{pairs}}|n_{\text{socks}} \sim {}^{n_{\text{socks}}}/_{2}\mathcal{B}e(15,2)$$

possible to

1. generate new values of n_{socks} and n_{pairs},
2. generate a new observation of X, number of unique socks out of 11.
3. accept the pair $(n_{\text{socks}}, n_{\text{pairs}})$ if the realisation of X is equal to 11



Meaning

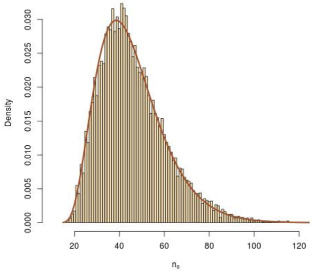


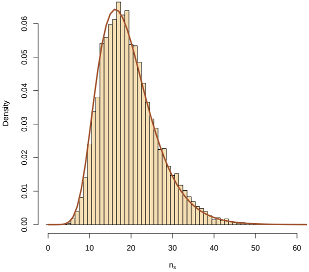
The outcome of this simulation method returns a distribution on the pair $(n_{\text{socks}}, n_{\text{pairs}})$ that is the conditional distribution of the pair given the observation $X = 11$

Proof: Generations from $\pi(n_{\text{socks}}, n_{\text{pairs}})$ are accepted with probability

$$\mathbb{P}\left\{X = 11 | (n_{\text{socks}}, n_{\text{pairs}})\right\}$$

Meaning





The outcome of this simulation method returns a distribution on the pair $(n_{\text{socks}}, n_{\text{pairs}})$ that is the conditional distribution of the pair given the observation $X = 11$

Proof: Hence accepted values distributed from

$$\pi(n_{\text{socks}}, n_{\text{pairs}}) \times \mathbb{P}\left\{X = 11 | (n_{\text{socks}}, n_{\text{pairs}})\right\} = \pi(n_{\text{socks}}, n_{\text{pairs}} | X = 11)$$

Econ'ections

1. simulation-based methods in Econometrics
2. Genetics of ABC
3. Approximate Bayesian computation
4. ABC for model choice
5. ABC model choice via random forests
6. ABC estimation via random forests
7. [some] asymptotics of ABC

Usages of simulation in Econometrics

Similar exploration of simulation-based techniques in Econometrics

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

[Gouriéroux & Monfort, 1996]

Simulated method of moments

Given observations $y^o_{1:n}$ from a model

$$y_t = r(y_{1:(t-1)}, \epsilon_t, \theta)\,, \quad \epsilon_t \sim g(\cdot)$$

simulate $\epsilon^\star_{1:n}$, derive

$$y^\star_t(\theta) = r(y_{1:(t-1)}, \epsilon^\star_t, \theta)$$

and estimate θ by

$$\arg\min_\theta \sum_{t=1}^n (y^o_t - y^\star_t(\theta))^2$$

Simulated method of moments

Given observations $y^o_{1:n}$ from a model

$$y_t = r(y_{1:(t-1)}, \epsilon_t, \theta)\,, \quad \epsilon_t \sim g(\cdot)$$

simulate $\epsilon^\star_{1:n}$, derive

$$y^\star_t(\theta) = r(y_{1:(t-1)}, \epsilon^\star_t, \theta)$$

and estimate θ by

$$\arg\min_\theta \left\{ \sum_{t=1}^n y^o_t - \sum_{t=1}^n y^\star_t(\theta) \right\}^2$$

Method of simulated moments

Given a statistic vector $K(y)$ with

$$\mathbb{E}_\theta[K(Y_t)|y_{1:(t-1)}] = k(y_{1:(t-1)};\theta)$$

find an *unbiased estimator* of $k(y_{1:(t-1)};\theta)$,

$$\tilde{k}(\epsilon_t, y_{1:(t-1)};\theta)$$

Estimate θ by

$$\arg\min_\theta \left|\left|\sum_{t=1}^{n} \left[K(y_t) - \sum_{s=1}^{S} \tilde{k}(\epsilon_t^s, y_{1:(t-1)};\theta)/S\right]\right|\right|$$

[Pakes & Pollard, 1989]

Indirect inference

Minimise (in θ) the distance between estimators $\hat{\beta}$ based on pseudo-models 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)

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

leading to

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

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)

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

leading to

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

when

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

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...*

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]

AR(2) vs. MA(1) example

true (AR) model

$$y_t = \epsilon_t - \theta \epsilon_{t-1}$$

and [wrong!] auxiliary (MA) model

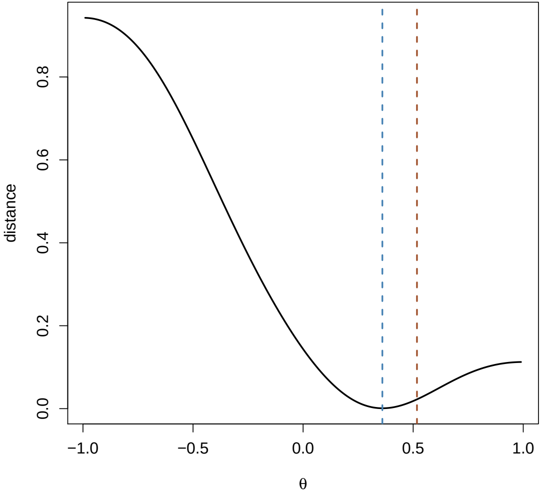
$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + u_t$$

R code

```
x=eps=rnorm(250)
x[2:250]=x[2:250]-0.5*x[1:249]
simeps=rnorm(250)
propeta=seq(-.99,.99,le=199)
dist=rep(0,199)
bethat=as.vector(arima(x,c(2,0,0),incl=FALSE)$coef)
for (t in 1:199)
  dist[t]=sum((as.vector(arima(c(simeps[1],simeps[2:250]-propeta[t]*
  simeps[1:249]),c(2,0,0),incl=FALSE)$coef)-bethat)^2)
```

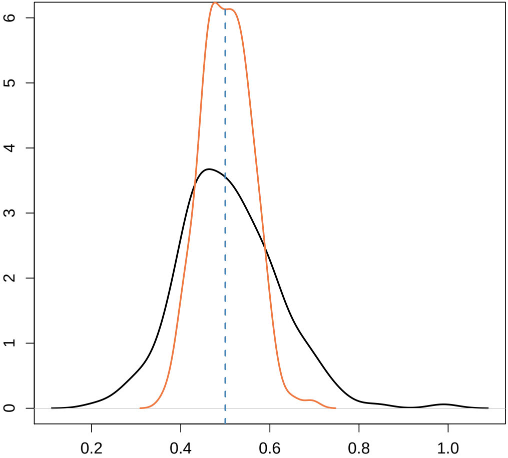
AR(2) vs. MA(1) example

One sample:



AR(2) vs. MA(1) example

Many samples:



Choice of pseudo-model

Pick model such that

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

[Frigessi & Heggland, 2004]

ABC using indirect inference (1)

> *We present a novel approach for developing summary statistics for use in approximate Bayesian computation (ABC) algorithms by using indirect inference(...) In the indirect inference approach to ABC the parameters of an auxiliary model fitted to the data become the summary statistics. Although applicable to any ABC technique, we embed this approach within a sequential Monte Carlo algorithm that is completely adaptive and requires very little tuning(...)*

[Drovandi, Pettitt & Faddy, 2011]

ⓒ Indirect inference provides summary statistics for ABC...

ABC using indirect inference (2)

> *...the above result shows that, in the limit as $h \to 0$, ABC will be more accurate than an indirect inference method whose auxiliary statistics are the same as the summary statistic that is used for ABC(...) Initial analysis showed that which method is more accurate depends on the true value of θ.*

[Fearnhead and Prangle, 2012]

ⓒ Indirect inference provides estimates rather than global inference...

Genetics of ABC

1. simulation-based methods in Econometrics
2. **Genetics of ABC**
3. Approximate Bayesian computation
4. ABC for model choice
5. ABC model choice via random forests
6. ABC estimation via random forests
7. [some] asymptotics of ABC

Genetic background of ABC

ABC is a recent computational technique that only requires a generative model, i.e., being able to sample from the density $f(\cdot|\boldsymbol{\theta})$

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]

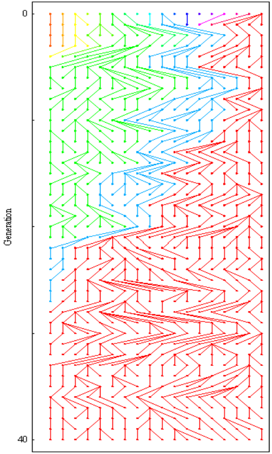
Population genetics

[Part derived from the teaching material of Raphael Leblois, ENS Lyon, November 2010]

- Describe the genotypes, estimate the alleles frequencies, determine their distribution among individuals, populations and between populations;
- Predict and understand the evolution of gene frequencies in populations as a result of various factors.

© Analyses the effect of various evolutive forces (mutation, drift, migration, selection) on the evolution of gene frequencies in time and space.

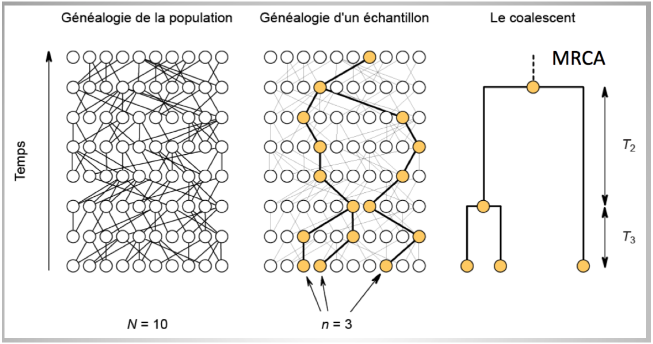
Wright-Fisher model



- A population of constant size, in which individuals reproduce at the same time.
- Each gene in a generation is a copy of a gene of the previous generation.
- In the absence of mutation and selection, allele frequencies derive inevitably until the fixation of an allele.

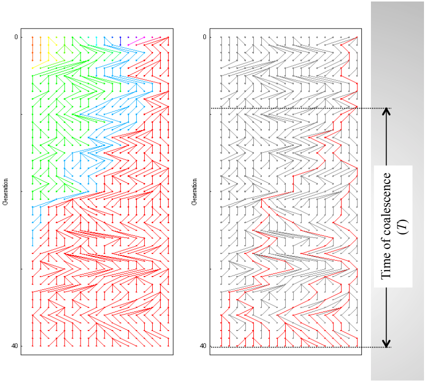
Coalescent theory

[Kingman, 1982; Tajima, Tavaré, &tc]



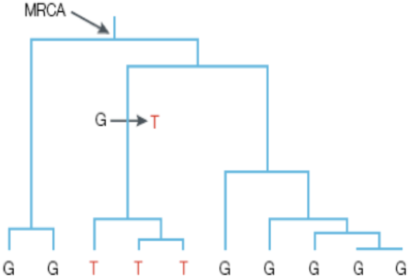
Coalescence theory interested in the genealogy of a sample of genes back in time to the common ancestor of the sample.

Common ancestor



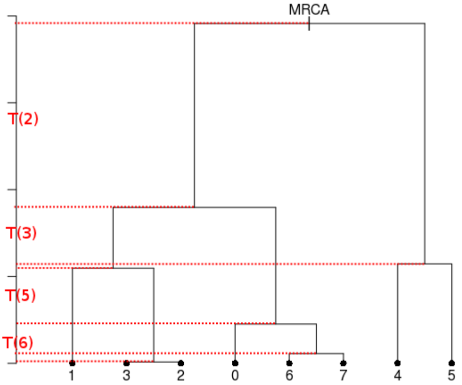
The different lineages merge when we go back in the past.

Neutral mutations



- Under the assumption of neutrality, the mutations are independent of the genealogy.
- We construct the genealogy according to the demographic parameters, then we add a posteriori the mutations.

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

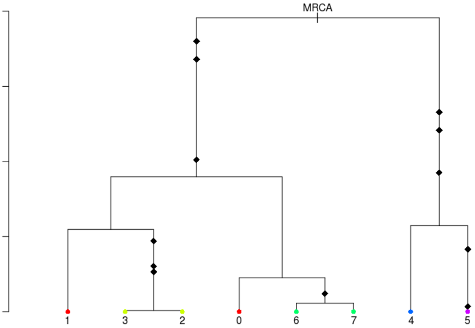


Kingman's genealogy

When time axis is normalized,

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

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



Kingman's genealogy

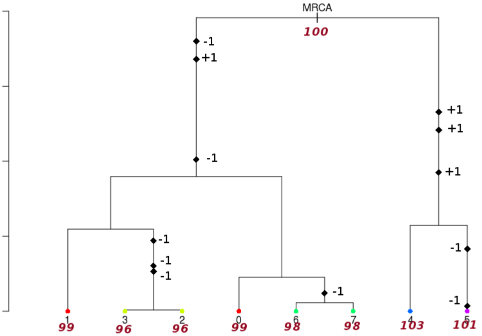
When time axis is normalized,

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

Mutations according to the Simple stepwise Mutation Model (SMM)

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

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. $1/2$

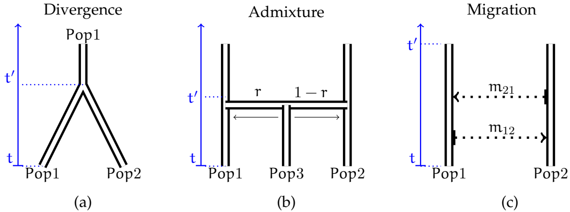
Much more interesting models. . .

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

Available population scenarios

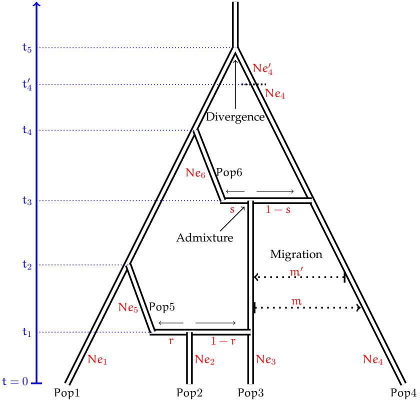
Between populations: three types of events, backward in time

- **the divergence** is the fusion between two populations,
- **the admixture** is the split of a population into two parts,
- **the migration** allows the move of some lineages of a population to another.



A complex scenario

The goal is to discriminate between different population scenarios from a dataset of polymorphism (DNA sample) **y** observed at the present time.



Demo-genetic inference

Each model is characterized by a set of parameters $\boldsymbol{\theta}$ 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 can not calculate the likelihood of the polymorphism data $f(\mathbf{y}|\boldsymbol{\theta})$.

Untractable likelihood

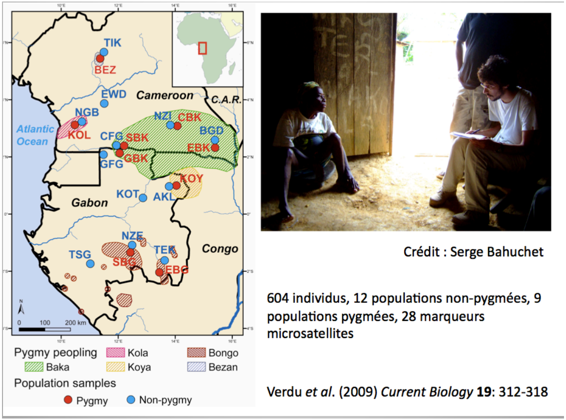
Missing (too missing!) data structure:

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

The genealogies are considered as nuisance parameters.

This problematic thus differs from the phylogenetic approach where the tree is the parameter of interesst.

A genuine example of application



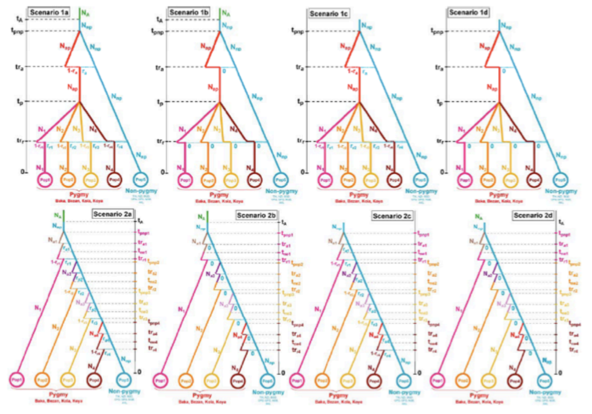
Crédit : Serge Bahuchet

604 individus, 12 populations non-pygmées, 9 populations pygmées, 28 marqueurs microsatellites

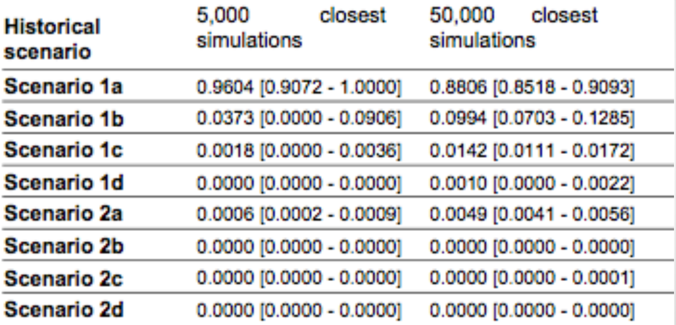
Verdu *et al.* (2009) *Current Biology* **19**: 312-318

Pygmies populations: do they have a common origin? Is there a lot of exchanges between pygmies and non-pygmies populations?

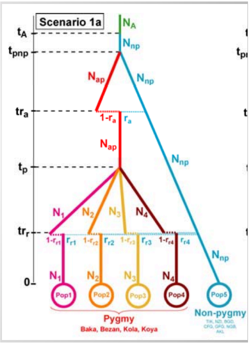
Scenarios under competition



Simulation results

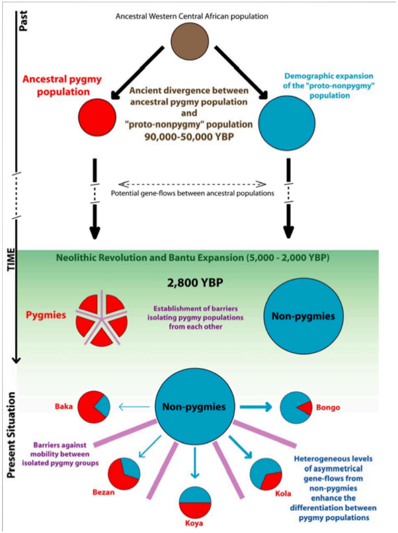


Historical scenario	5,000 closest simulations	50,000 closest simulations
Scenario 1a	0.9604 [0.9072 - 1.0000]	0.8806 [0.8518 - 0.9093]
Scenario 1b	0.0373 [0.0000 - 0.0906]	0.0994 [0.0703 - 0.1285]
Scenario 1c	0.0018 [0.0000 - 0.0036]	0.0142 [0.0111 - 0.0172]
Scenario 1d	0.0000 [0.0000 - 0.0000]	0.0010 [0.0000 - 0.0022]
Scenario 2a	0.0006 [0.0002 - 0.0009]	0.0049 [0.0041 - 0.0056]
Scenario 2b	0.0000 [0.0000 - 0.0000]	0.0000 [0.0000 - 0.0000]
Scenario 2c	0.0000 [0.0000 - 0.0000]	0.0000 [0.0000 - 0.0001]
Scenario 2d	0.0000 [0.0000 - 0.0000]	0.0000 [0.0000 - 0.0000]



ⓒ Scenario 1A is chosen.

Most likely scenario



Instance of ecological questions [message in a beetle]

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

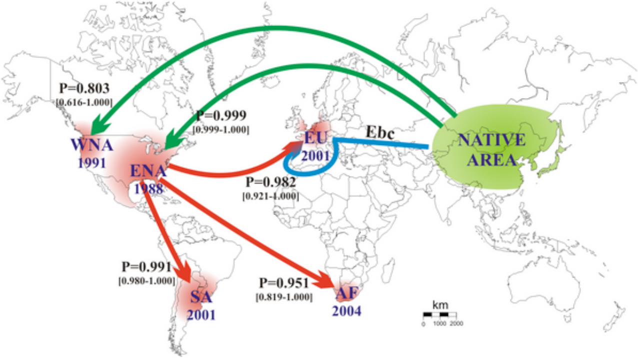
- Why did the chicken cross the road?


dreamstime.com

[Lombaert & al., 2010, PLoS ONE]

▸ beetles in forests

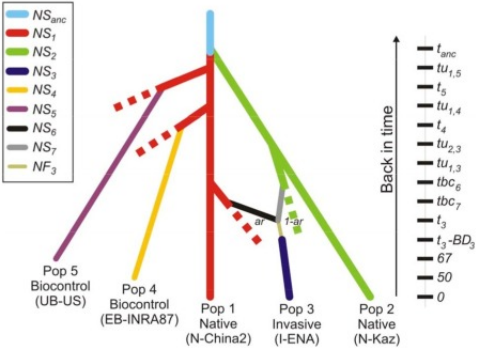
Worldwide invasion routes of *Harmonia Axyridis*



For each outbreak, the arrow indicates the most likely invasion pathway and the associated posterior probability, with 95% credible intervals in brackets

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

Worldwide invasion routes of *Harmonia Axyridis*



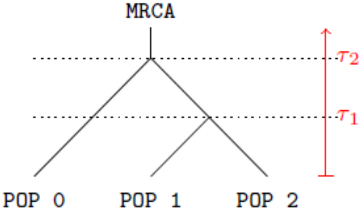
For each outbreak, the arrow indicates the most likely invasion pathway and the associated posterior probability, with 95% credible intervals in brackets

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

A population genetic illustration of ABC model choice

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 of ABC model choice

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 of ABC model choice

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 of ABC model choice

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

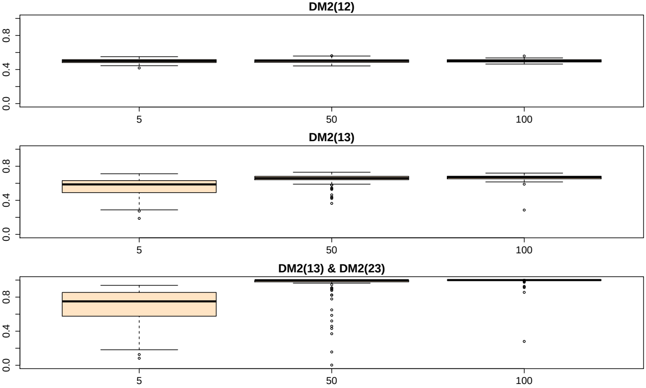
	Model 1	Model 2
$\mathbb{E}\left\{(\delta_\mu)^2_{1,2}\right\}$	$2\mu_1 t'$	$2\mu_2 t'$
$\mathbb{E}\left\{(\delta_\mu)^2_{1,3}\right\}$	$2\mu_1 t$	$2\mu_2 t'$
$\mathbb{E}\left\{(\delta_\mu)^2_{2,3}\right\}$	$2\mu_1 t'$	$2\mu_2 t$

A population genetic illustration of ABC model choice

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 of ABC model choice



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

Untractable likelihoods

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

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

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

Untractable likelihoods

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

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

is impossible or too costly because of the dimension of $\mathbf{z}$

© MCMC cannot be implemented!

Illustration

Example (Ising & Potts models)

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

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

where $l\sim i$ denotes a neighbourhood relation, n moderately large prohibits the computation of the normalising constant $\mathfrak{Z}_\theta$

Illustration

Example (Ising & Potts models)

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

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

where $l\sim i$ denotes a neighbourhood relation, n moderately large prohibits the computation of the normalising constant $\mathfrak{Z}_\theta$

Special case of the intractable normalising constant, making the likelihood impossible to compute

The ABC method

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

The ABC method

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

The ABC method

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

ABC algorithm

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

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

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

[Tavaré et al., 1997]

Why does it work?!

The proof is trivial:

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

[Accept–Reject 101]

Earlier occurrence

> *'Bayesian statistics and Monte Carlo methods are ideally suited to the task of passing many models over one dataset'*
>
> [Don Rubin, *Annals of Statistics*, 1984]

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

A as A...pproximative

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

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

where ϱ is a distance

A as A...pproximative

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

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

where ϱ is a distance
Output distributed from

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

[Pritchard et al., 1999]

ABC algorithm

Algorithm 1 Likelihood-free rejection sampler 2

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

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

Output

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

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

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

Output

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

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

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

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

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

Convergence of ABC (first attempt)

What happens when $\epsilon \to 0$?

Convergence of ABC (first attempt)

What happens when $\epsilon \to 0$?

If $f(\cdot|\theta)$ is continuous in y, uniformly in θ [!], given an arbitrary $\delta > 0$, there exists ϵ_0 such that $\epsilon < \epsilon_0$ implies

$$\frac{\pi(\theta) \int f(\mathbf{z}|\theta)\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z})\,\mathrm{d}\mathbf{z}}{\int_{A_{\epsilon,\mathbf{y}}\times\Theta} \pi(\theta)f(\mathbf{z}|\theta)\mathrm{d}\mathbf{z}\mathrm{d}\theta} \in \frac{\pi(\theta)f(\mathbf{y}|\theta)(1\mp\delta)\mu(\mathfrak{B}_\epsilon)}{\int_\Theta \pi(\theta)f(\mathbf{y}|\theta)\mathrm{d}\theta(1\pm\delta)\mu(\mathfrak{B}_\epsilon)}$$

Convergence of ABC (first attempt)

What happens when $\epsilon \to 0$?

If $f(\cdot|\theta)$ is continuous in y, uniformly in θ [!], given an arbitrary $\delta > 0$, there exists ϵ_0 such that $\epsilon < \epsilon_0$ implies

$$\frac{\pi(\theta) \int f(\mathbf{z}|\theta)\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z})\,\mathrm{d}\mathbf{z}}{\int_{A_{\epsilon,\mathbf{y}}\times\Theta} \pi(\theta)f(\mathbf{z}|\theta)\mathrm{d}\mathbf{z}\mathrm{d}\theta} \in \frac{\pi(\theta)f(\mathbf{y}|\theta)(1\mp\delta)\xcancel{\mu(\mathfrak{B}_\epsilon)}}{\int_\Theta \pi(\theta)f(\mathbf{y}|\theta)\mathrm{d}\theta(1\pm\delta)\xcancel{\mu(\mathfrak{B}_\epsilon)}}$$

Convergence of ABC (first attempt)

What happens when $\epsilon \to 0$?

If $f(\cdot|\theta)$ is continuous in y, uniformly in θ [!], given an arbitrary $\delta > 0$, there exists ϵ_0 such that $\epsilon < \epsilon_0$ implies

$$\frac{\pi(\theta)\int f(\mathbf{z}|\theta)\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z})\,\mathrm{d}\mathbf{z}}{\int_{A_{\epsilon,\mathbf{y}}\times\Theta}\pi(\theta)f(\mathbf{z}|\theta)\mathrm{d}\mathbf{z}\mathrm{d}\theta} \in \frac{\pi(\theta)f(\mathbf{y}|\theta)(1\mp\delta)\cancel{\mu(\mathfrak{B}_\epsilon)}}{\int_\Theta \pi(\theta)f(\mathbf{y}|\theta)\mathrm{d}\theta(1\pm\delta)\cancel{\mu(\mathfrak{B}_\epsilon)}}$$

[Proof extends to other continuous-in-0 kernels K_ϵ]

Convergence of ABC (second attempt)

What happens when $\epsilon \to 0$?

Convergence of ABC (second attempt)

What happens when $\epsilon \to 0$?

For $B \subset \Theta$, we have

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

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

Probit modelling on Pima Indian women

Example (R benchmark)

200 Pima Indian women with observed variables

- plasma glucose concentration in oral glucose tolerance test
- diastolic blood pressure
- diabetes pedigree function
- presence/absence of diabetes

Probit modelling on Pima Indian women

Example (R benchmark)

200 Pima Indian women with observed variables

- plasma glucose concentration in oral glucose tolerance test
- diastolic blood pressure
- diabetes pedigree function
- presence/absence of diabetes

Probability of diabetes function of above variables

$$\mathbb{P}(y = 1|x) = \Phi(x_1\beta_1 + x_2\beta_2 + x_3\beta_3)\,,$$

Probit modelling on Pima Indian women

Example (R benchmark)

200 Pima Indian women with observed variables

- plasma glucose concentration in oral glucose tolerance test
- diastolic blood pressure
- diabetes pedigree function
- presence/absence of diabetes

Probability of diabetes function of above variables

$$\mathbb{P}(y = 1|x) = \Phi(x_1\beta_1 + x_2\beta_2 + x_3\beta_3)\,,$$

Test of $H_0 : \beta_3 = 0$ for 200 observations of Pima.tr based on a g-prior modelling:

$$\beta \sim \mathcal{N}_3(0, n\left(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$$

Probit modelling on Pima Indian women

Example (R benchmark)

200 Pima Indian women with observed variables

- plasma glucose concentration in oral glucose tolerance test
- diastolic blood pressure
- diabetes pedigree function
- presence/absence of diabetes

Probability of diabetes function of above variables

$$\mathbb{P}(y=1|x) = \Phi(x_1\beta_1 + x_2\beta_2 + x_3\beta_3)\,,$$

Test of $H_0 : \beta_3 = 0$ for 200 observations of Pima.tr based on a g-prior modelling:

$$\beta \sim \mathcal{N}_3(0, n\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\Big)$$

Use of importance function inspired from the MLE estimate distribution

$$\beta \sim \mathcal{N}(\hat\beta, \hat\Sigma)$$

Pima Indian benchmark

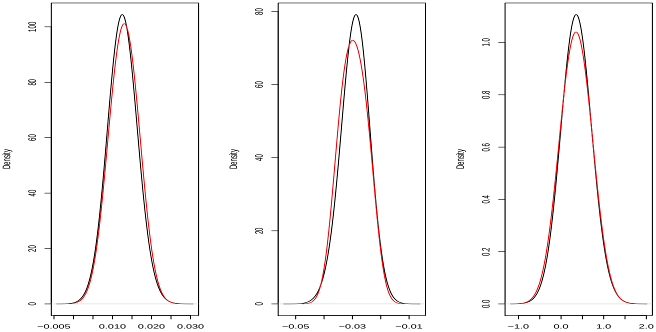


Figure: Comparison between density estimates of the marginals on β_1 (left), β_2 (center) and β_3 (right) from ABC rejection samples (red) and MCMC samples (black)

.

MA example

Back to the MA(q) model

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

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

Back to the MA(q) model

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

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

MA example (2)

ABC algorithm thus made of

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

MA example (2)

ABC algorithm thus made of

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

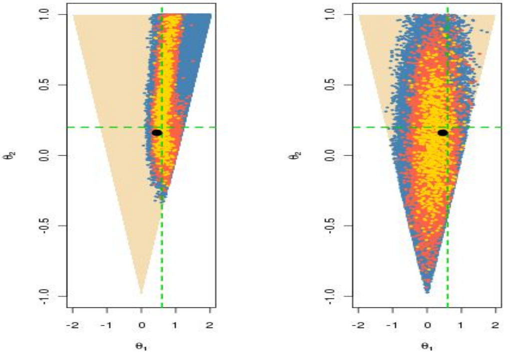
Distance: basic distance between the series

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

or distance between summary statistics like the q autocorrelations

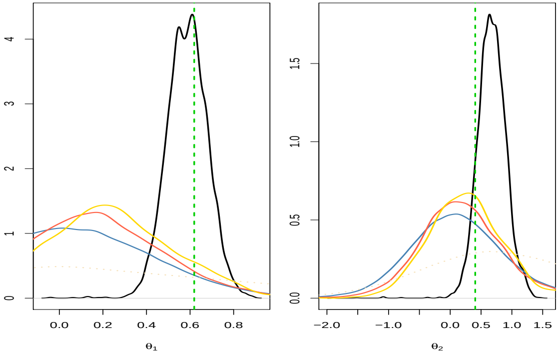
$$\tau_j = \sum_{t=j+1}^{T} x_t x_{t-j}$$

Comparison of distance impact



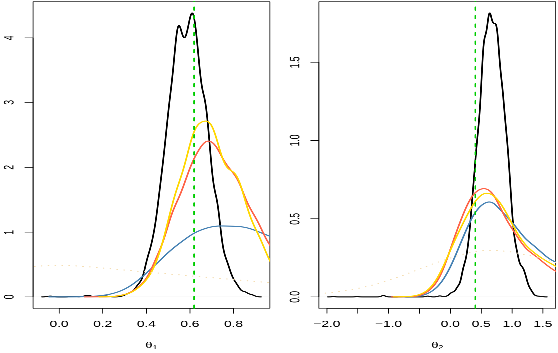
Evaluation of the tolerance on the ABC sample against both distances ($\epsilon = 100\%, 10\%, 1\%, 0.1\%$) for an MA(2) model

Comparison of distance impact



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

Comparison of distance impact



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

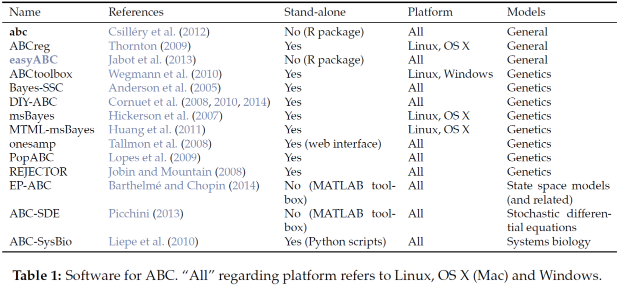
Homonomy

The ABC algorithm is not to be confused with the ABC algorithm

> *The Artificial Bee Colony algorithm is a swarm based meta-heuristic algorithm that was introduced by Karaboga in 2005 for optimizing numerical problems. It was inspired by the intelligent foraging behavior of honey bees. The algorithm is specifically based on the model proposed by Tereshko and Loengarov (2005) for the foraging behaviour of honey bee colonies. The model consists of three essential components: employed and unemployed foraging bees, and food sources. The first two components, employed and unemployed foraging bees, search for rich food sources (...) close to their hive. The model also defines two leading modes of behaviour (...): recruitment of foragers to rich food sources resulting in positive feedback and abandonment of poor sources by foragers causing negative feedback.*

[Karaboga, Scholarpedia]

ABC R packages



Name	References	Stand-alone	Platform	Models
abc	Csilléry et al. (2012)	No (R package)	All	General
ABCreg	Thornton (2009)	Yes	Linux, OS X	General
easyABC	Jabot et al. (2013)	No (R package)	All	General
ABCtoolbox	Wegmann et al. (2010)	Yes	Linux, Windows	Genetics
Bayes-SSC	Anderson et al. (2005)	Yes	All	Genetics
DIY-ABC	Cornuet et al. (2008, 2010, 2014)	Yes	All	Genetics
msBayes	Hickerson et al. (2007)	Yes	Linux, OS X	Genetics
MTML-msBayes	Huang et al. (2011)	Yes	Linux, OS X	Genetics
onesamp	Tallmon et al. (2008)	Yes (web interface)	All	Genetics
PopABC	Lopes et al. (2009)	Yes	All	Genetics
REJECTOR	Jobin and Mountain (2008)	Yes	All	Genetics
EP-ABC	Barthelmé and Chopin (2014)	No (MATLAB toolbox)	All	State space models (and related)
ABC-SDE	Picchini (2013)	No (MATLAB toolbox)	All	Stochastic differential equations
ABC-SysBio	Liepe et al. (2010)	Yes (Python scripts)	All	Systems biology

Table 1: Software for ABC. "All" regarding platform refers to Linux, OS X (Mac) and Windows.

- abc version 2.1

  [Csilléry et al., 2015]

- abctools version 1.1.1

  [Nunes & Prangle, 2017]

- abcrf version 1.7

  [Marin et al., 2017]

ABC advances

Simulating from the prior is often poor in efficiency

ABC advances

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

[Marjoram et al, 2003; Bortot et al., 2007, Sisson et al., 2007]

ABC advances

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

[Marjoram et al, 2003; 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]

ABC advances

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

[Marjoram et al, 2003; 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$_\mu$]

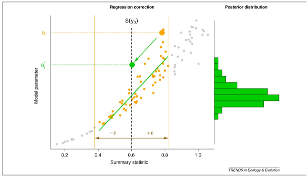
[Ratmann et al., 2009]

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

▸ (use with BIC)

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



[Csilléry et al., TEE, 2010]

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

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

[Beaumont et al., 2002, Genetics]

ABC-NP (regression)

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

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

or

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

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

ABC-NP (regression)

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

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

or

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

[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

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

leads to the NP estimate of the posterior conditional density

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

[Blum, JASA, 2010]

ABC-NP (density estimations)

Other versions incorporating regression adjustments

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

ABC-NP (density estimations)

Other versions incorporating regression adjustments

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

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

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

In all cases, error

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

[standard NP calculations]

ABC-NCH

Incorporating non-linearities and heterocedasticities:

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

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 (2)

Why neural network?

ABC-NCH (2)

Why neural network?

- fights curse of dimensionality
- selects relevant summary statistics
- provides automated dimension reduction
- offers a model choice capability
- improves upon multinomial logistic

[Blum & François, 2009]

ABC as knn

[Biau et al., 2013, Annales de l'IHP]

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

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

ABC as knn

[Biau et al., 2013, Annales de l'IHP]

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 as knn

[Biau et al., 2013, Annales de l'IHP]

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]

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. (2013) 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., 2013]

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., 2013]

Drag: Only applies to sufficient summary statistics

How Bayesian is ABC..?

- may be a convergent method of inference (meaningful? sufficient? foreign?)
- approximation error unknown (w/o massive simulation)
- pragmatic/empirical B (there is no other solution!)
- many calibration issues (tolerance, distance, statistics)
- the NP side should be incorporated into the whole B picture
- the approximation error should also be part of the B inference

ABC-MCMC

Markov chain $(\theta^{(t)})$ created via the transition function

$$\theta^{(t+1)} = \begin{cases} \theta' \sim K_\omega(\theta'|\theta^{(t)}) & \text{if } x \sim f(x|\theta') \text{ is such that } x = y \\ & \quad \text{and } u \sim \mathcal{U}(0,1) \leq \frac{\pi(\theta')K_\omega(\theta^{(t)}|\theta')}{\pi(\theta^{(t)})K_\omega(\theta'|\theta^{(t)})}\,, \\ \theta^{(t)} & \text{otherwise,} \end{cases}$$

ABC-MCMC

Markov chain $(\theta^{(t)})$ created via the transition function

$$\theta^{(t+1)} = \begin{cases} \theta' \sim K_\omega(\theta'|\theta^{(t)}) & \text{if } x \sim f(x|\theta') \text{ is such that } x = y \\ & \quad \text{and } u \sim \mathcal{U}(0,1) \leq \frac{\pi(\theta')K_\omega(\theta^{(t)}|\theta')}{\pi(\theta^{(t)})K_\omega(\theta'|\theta^{(t)})}\,, \\ \theta^{(t)} & \text{otherwise,} \end{cases}$$

has the posterior $\pi(\theta|y)$ as stationary distribution

[Marjoram et al, 2003]

ABC-MCMC (2)

Algorithm 2 Likelihood-free MCMC sampler

```
Use Algorithm 1 to get $(\theta^{(0)}, \mathbf{z}^{(0)})$
for $t = 1$ to $N$ do
    Generate $\theta'$ from $K_\omega\left(\cdot|\theta^{(t-1)}\right)$,
    Generate $\mathbf{z}'$ from the likelihood $f(\cdot|\theta')$,
    Generate $u$ from $\mathcal{U}_{[0,1]}$,
    if $u \leq \frac{\pi(\theta')K_\omega(\theta^{(t-1)}|\theta')}{\pi(\theta^{(t-1)}K_\omega(\theta'|\theta^{(t-1)})}\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}')$ then
        set $(\theta^{(t)}, \mathbf{z}^{(t)}) = (\theta', \mathbf{z}')$
    else
        $(\theta^{(t)}, \mathbf{z}^{(t)})) = (\theta^{(t-1)}, \mathbf{z}^{(t-1)})$,
    end if
end for
```

Why does it work?

Acceptance probability does not involve calculating the likelihood and

$$
\begin{aligned}
\frac{\pi_\epsilon(\boldsymbol{\theta}', \mathbf{z}'|\mathbf{y})}{\pi_\epsilon(\boldsymbol{\theta}^{(t-1)}, \mathbf{z}^{(t-1)}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})f(\mathbf{z}'|\boldsymbol{\theta}')}
&= \frac{\pi(\boldsymbol{\theta}')\ \cancel{f(\mathbf{z}'|\boldsymbol{\theta}')}\ \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}')}{\pi(\boldsymbol{\theta}^{(t-1)})\, f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})\, \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}^{(t-1)})} \\
&\times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')\ f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})\ \cancel{f(\mathbf{z}'|\boldsymbol{\theta}')}}
\end{aligned}
$$

Why does it work?

Acceptance probability does not involve calculating the likelihood and

$$\begin{aligned}
&\frac{\pi_\epsilon(\boldsymbol{\theta}', \mathbf{z}'|\mathbf{y})}{\pi_\epsilon(\boldsymbol{\theta}^{(t-1)}, \mathbf{z}^{(t-1)}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})f(\mathbf{z}'|\boldsymbol{\theta}')} \\
&\qquad = \frac{\pi(\boldsymbol{\theta}')\ \cancel{f(\mathbf{z}'|\boldsymbol{\theta}')}\ \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}')}{\pi(\boldsymbol{\theta}^{(t-1)})\ \cancel{f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}\ \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}^{(t-1)})} \\
&\qquad \times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')\ \cancel{f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})\ \cancel{f(\mathbf{z}'|\boldsymbol{\theta}')}}
\end{aligned}$$

Why does it work?

Acceptance probability does not involve calculating the likelihood and

$$\begin{aligned}
\frac{\pi_\epsilon(\boldsymbol{\theta}', \mathbf{z}'|\mathbf{y})}{\pi_\epsilon(\boldsymbol{\theta}^{(t-1)}, \mathbf{z}^{(t-1)}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})f(\mathbf{z}'|\boldsymbol{\theta}')}
&= \frac{\pi(\boldsymbol{\theta}')\ \cancel{f(\mathbf{z}'|\boldsymbol{\theta}')}\ \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}')}{\pi(\boldsymbol{\theta}^{(t-1)})\ \cancel{f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}\ \cancel{\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}^{(t-1)})}} \\
&\times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')\ \cancel{f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})\ \cancel{f(\mathbf{z}'|\boldsymbol{\theta}')}} \\
&= \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t-1)}q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})}\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z}')
\end{aligned}$$

ABC$_\mu$

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

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 **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 **y** when $\mathbf{z} \sim f(\mathbf{z}|\theta)$

ABC$_\mu$

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

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.

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$_\mu$ details

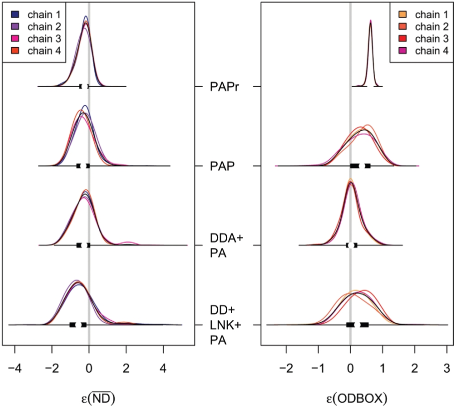
Multidimensional distances ρ_k $(k = 1, \ldots, K)$ and errors $\epsilon_k = \rho_k(\eta_k(\mathbf{z}), \eta_k(\mathbf{y}))$, with

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

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

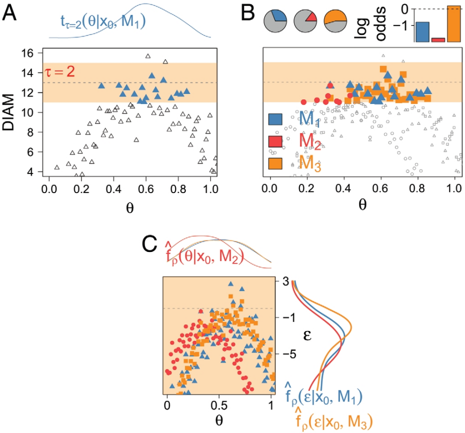
$$\frac{\pi(\theta', \epsilon')}{\pi(\theta, \epsilon)} \frac{q(\theta', \theta) q(\epsilon', \epsilon)}{q(\theta, \theta') q(\epsilon, \epsilon')} \frac{\min_k \hat{\xi}_k(\epsilon'|\mathbf{y}, \theta')}{\min_k \hat{\xi}_k(\epsilon|\mathbf{y}, \theta)}$$

ABC$_\mu$ multiple errors



[© Ratmann et al., PNAS, 2009]

ABC$_\mu$ for model choice



[© Ratmann et al., PNAS, 2009]

Questions about ABC$_\mu$

For each model under comparison, marginal posterior on ϵ used to assess the fit of the model (HPD includes 0 or not).

Questions about ABC$_\mu$

For each model under comparison, marginal posterior on ϵ used to assess the fit of the model (HPD includes 0 or not).

- Is the data informative about ϵ? [Identifiability]
- How is the prior $\pi(\epsilon)$ impacting the comparison?
- How is using both $\xi(\epsilon|x_0, \theta)$ and $\pi_\epsilon(\epsilon)$ compatible with a standard probability model? [remindful of ▸ Wilkinson]
- Where is the penalisation for complexity in the model comparison?

[X, Mengersen & Chen, 2010, PNAS]

A PMC version

Use of the same kernel idea as ABC-PRC (Sisson et al., 2007) but with IS correction
Generate a sample at iteration t by

$$\hat{\pi}_t(\theta^{(t)}) \propto \sum_{j=1}^{N} \omega_j^{(t-1)} K_t(\theta^{(t)}|\theta_j^{(t-1)})$$

modulo acceptance of the associated x_t, and use an importance weight associated with an accepted simulation $\theta_i^{(t)}$

$$\omega_i^{(t)} \propto \pi(\theta_i^{(t)}) \big/ \hat{\pi}_t(\theta_i^{(t)}) \,.$$

ⓒ **Still likelihood free**

[Beaumont et al., 2009]

ABC-PMC algorithm

Given a decreasing sequence of approximation levels $\epsilon_1 \geq \ldots \geq \epsilon_T$,

```
1. At iteration $t = 1$,
        For $i = 1, ..., N$
            Simulate $\theta_i^{(1)} \sim \pi(\theta)$ and $x \sim f(x|\theta_i^{(1)})$ until $\varrho(x, y) < \epsilon_1$
            Set $\omega_i^{(1)} = 1/N$
   Take $\tau^2$ as twice the empirical variance of the $\theta_i^{(1)}$'s
2. At iteration $2 \leq t \leq T$,
        For $i = 1, ..., N$, repeat
            Pick $\theta_i^\star$ from the $\theta_j^{(t-1)}$'s with probabilities $\omega_j^{(t-1)}$
            generate $\theta_i^{(t)}|\theta_i^\star \sim \mathcal{N}(\theta_i^\star, \sigma_t^2)$ and $x \sim f(x|\theta_i^{(t)})$
        until $\varrho(x, y) < \epsilon_t$
        Set $\omega_i^{(t)} \propto \pi(\theta_i^{(t)}) / \sum_{j=1}^N \omega_j^{(t-1)} \varphi\left(\sigma_t^{-1}\left\{\theta_i^{(t)} - \theta_j^{(t-1)}\right\}\right)$
   Take $\tau_{t+1}^2$ as twice the weighted empirical variance of the $\theta_i^{(t)}$'s
```

Sequential Monte Carlo

SMC is a simulation technique that approximates a sequence of related probability distributions π_n with π_0 "easy" and π_T as target.

Iterated IS as PMC: particles moved from time n to time n via kernel K_n and use of a sequence of extended targets $\tilde{\pi}_n$

$$\tilde{\pi}_n(\mathbf{z}_{0:n}) = \pi_n(z_n) \prod_{j=0}^{n} L_j(z_{j+1}, z_j)$$

where the L_j's are backward Markov kernels [check that $\pi_n(z_n)$ is a marginal]

[Del Moral, Doucet & Jasra, Series B, 2006]

Sequential Monte Carlo (2)

Algorithm 3 SMC sampler

```
sample $z_i^{(0)} \sim \gamma_0(x)$ $(i = 1, \ldots, N)$
compute weights $w_i^{(0)} = \pi_0(z_i^{(0)})/\gamma_0(z_i^{(0)})$
for $t = 1$ to $N$ do
    if $\mathrm{ESS}(w^{(t-1)}) < N_T$ then
        resample $N$ particles $z^{(t-1)}$ and set weights to 1
    end if
    generate $z_i^{(t-1)} \sim K_t(z_i^{(t-1)}, \cdot)$ and set weights to
        $$w_i^{(t)} = w_{i-1}^{(t-1)} \frac{\pi_t(z_i^{(t)}))L_{t-1}(z_i^{(t)}), z_i^{(t-1)}))}{\pi_{t-1}(z_i^{(t-1)}))K_t(z_i^{(t-1)}), z_i^{(t)}))}$$
end for
```

[Del Moral, Doucet & Jasra, Series B, 2006]

ABC-SMC

[Del Moral, Doucet & Jasra, 2009]

True derivation of an SMC-ABC algorithm
Use of a kernel K_n associated with target π_{ϵ_n} and derivation of the backward kernel

$$L_{n-1}(z, z') = \frac{\pi_{\epsilon_n}(z')K_n(z', z)}{\pi_n(z)}$$

Update of the weights

$$w_{in} \propto w_{i(n-1)} \, \frac{\sum_{m=1}^{M} \mathbb{I}_{A_{\epsilon_n}}(x_{in}^m)}{\sum_{m=1}^{M} \mathbb{I}_{A_{\epsilon_{n-1}}}(x_{i(n-1)}^m)}$$

when $x_{in}^m \sim K(x_{i(n-1)}, \cdot)$

ABC-SMC$_M$

Modification: Makes M repeated simulations of the pseudo-data $\mathbf{z}$ given the parameter, rather than using a single $[M = 1]$ simulation, leading to weight that is proportional to the number of accepted $\mathbf{z}_i$s

$$\omega(\theta) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}_{\rho(\eta(\mathbf{y}),\eta(\mathbf{z}_i))<\epsilon}$$

[limit in M means exact simulation from (tempered) target]

Properties of ABC-SMC

The ABC-SMC method properly uses a backward kernel $L(z, z')$ to simplify the importance weight and to remove the dependence on the unknown likelihood from this weight. Update of importance weights is reduced to the ratio of the proportions of surviving particles

Major assumption: the forward kernel K is supposed to be invariant against the true target [tempered version of the true posterior]

Properties of ABC-SMC

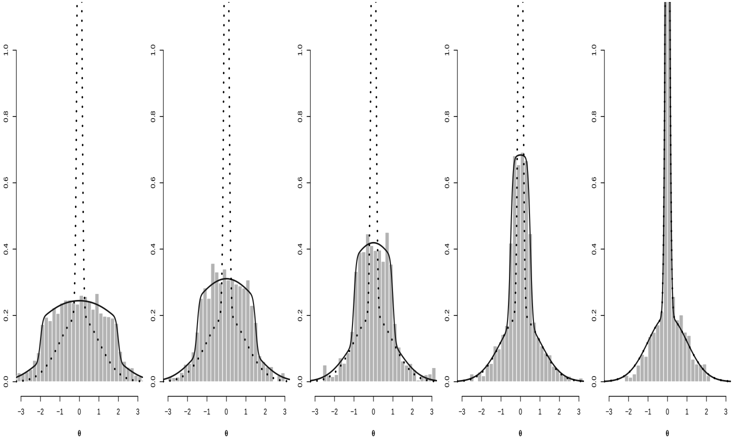
The ABC-SMC method properly uses a backward kernel $L(z, z')$ to simplify the importance weight and to remove the dependence on the unknown likelihood from this weight. Update of importance weights is reduced to the ratio of the proportions of surviving particles

Major assumption: the forward kernel K is supposed to be invariant against the true target [tempered version of the true posterior]

Adaptivity in ABC-SMC algorithm only found in on-line construction of the thresholds ϵ_t, slowly enough to keep a large number of accepted transitions

A mixture example (2)

Recovery of the target, whether using a fixed standard deviation of $\tau = 0.15$ or $\tau = 1/0.15$, or a sequence of adaptive τ_t's.



Wilkinson's exact BC

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]

Wilkinson's exact BC

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, 2013]

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_μ and assuming $\mathbf{y}$ 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

Noisy ABC

Idea: Modify the data from the start

$$\tilde{y} = y0 + \epsilon\zeta_1$$

with the same scale ϵ as ABC

[▸ see Fearnhead-Prangle]

run ABC on $\tilde{y}$

Noisy ABC

Idea: Modify the data from the start

$$\tilde{y} = y0 + \epsilon\zeta_1$$

with the same scale ϵ as ABC

[▸ see Fearnhead-Prangle]

run ABC on $\tilde{y}$
Then ABC produces an exact simulation from $\pi(\theta|\tilde{y}) = \pi(\theta|\tilde{y})$
[Dean et al., 2011; Fearnhead and Prangle, 2012]

Consistent noisy ABC

- 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]

Semi-automatic ABC

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

ABC then considered from a purely inferential viewpoint and calibrated for estimation purposes

Use of a 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

Semi-automatic ABC

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

ABC then considered from a purely inferential viewpoint and calibrated for estimation purposes

Use of a 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 [calibration constraint: ABC approximation with same posterior mean as the true randomised posterior]

Summary statistics

- Optimality of the posterior expectation $\mathbb{E}[\theta|\mathbf{y}]$ of the parameter of interest as summary statistics $\eta(\mathbf{y})$!

Summary 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

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

▸ bare summary

Details on Fearnhead and Prangle (F&P) ABC

Use of a summary statistic $S(\cdot)$, an importance proposal $g(\cdot)$, a kernel $K(\cdot) \leq 1$ and a bandwidth $h > 0$ such that

$$(\theta, \mathbf{y}_{\text{sim}}) \sim g(\theta) f(\mathbf{y}_{\text{sim}} | \theta)$$

is accepted with probability (hence the bound)

$$K[\{S(\mathbf{y}_{\text{sim}}) - \mathbf{s}_{\text{obs}}\}/h]$$

and the corresponding importance weight defined by

$$\pi(\theta) \big/ g(\theta)$$

[Fearnhead & Prangle, 2012]

Errors, errors, and errors

Three levels of approximation

- $\pi(\theta|\mathbf{y}_{\text{obs}})$ by $\pi(\theta|\mathbf{s}_{\text{obs}})$ loss of information

  [ignored]

- $\pi(\theta|\mathbf{s}_{\text{obs}})$ by

  $$\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) = \frac{\int \pi(\mathbf{s})K[\{\mathbf{s}-\mathbf{s}_{\text{obs}}\}/h]\pi(\theta|\mathbf{s})\,\text{d}\mathbf{s}}{\int \pi(\mathbf{s})K[\{\mathbf{s}-\mathbf{s}_{\text{obs}}\}/h]\,\text{d}\mathbf{s}}$$

  noisy observations

- $\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}})$ by importance Monte Carlo based on N simulations, represented by $\text{var}(a(\theta)|\mathbf{s}_{\text{obs}})/N_{\text{acc}}$ [expected number of acceptances]

[M. Twain/B. Disraeli]

Average acceptance asymptotics

For the average acceptance probability/approximate likelihood

$$p(\theta|\mathbf{s}_{\text{obs}}) = \int f(\mathbf{y}_{\text{sim}}|\theta)\, K[\{S(\mathbf{y}_{\text{sim}}) - \mathbf{s}_{\text{obs}}\}/h]\, \mathrm{d}\mathbf{y}_{\text{sim}}\,,$$

overall acceptance probability

$$p(\mathbf{s}_{\text{obs}}) = \int p(\theta|\mathbf{s}_{\text{obs}})\, \pi(\theta)\, \mathrm{d}\theta = \pi(\mathbf{s}_{\text{obs}})h^d + o(h^d)$$

[F&P, Lemma 1]

Optimal importance proposal

Best choice of importance proposal in terms of effective sample size

$$g^{\star}(\theta|\mathbf{s}_{\text{obs}}) \propto \pi(\theta) p(\theta|\mathbf{s}_{\text{obs}})^{1/2}$$

[Not particularly useful in practice]

Optimal importance proposal

Best choice of importance proposal in terms of effective sample size

$$g^\star(\theta|\mathbf{s}_{\text{obs}}) \propto \pi(\theta) p(\theta|\mathbf{s}_{\text{obs}})^{1/2}$$

[Not particularly useful in practice]

- note that $p(\theta|\mathbf{s}_{\text{obs}})$ is an approximate likelihood
- reminiscent of parallel tempering
- could be approximately achieved by attrition of half of the data

Calibration of *h*

> *"This result gives insight into how $S(\cdot)$ and h affect the Monte Carlo error. To minimize Monte Carlo error, we need h^d to be not too small. Thus ideally we want $S(\cdot)$ to be a low dimensional summary of the data that is sufficiently informative about θ that $\pi(\theta|\mathbf{s}_{obs})$ is close, in some sense, to $\pi(\theta|\mathbf{y}_{obs})$" (F&P, p.5)*

- turns h into an absolute value while it should be context-dependent and user-calibrated
- only addresses one term in the approximation error and acceptance probability ("curse of dimensionality")
- h large prevents $\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}})$ to be close to $\pi(\theta|\mathbf{s}_{\text{obs}})$
- d small prevents $\pi(\theta|\mathbf{s}_{\text{obs}})$ to be close to $\pi(\theta|\mathbf{y}_{\text{obs}})$ ("curse of [dis]information")

Calibrating ABC

"If π_{ABC} is calibrated, then this means that probability statements that are derived from it are appropriate, and in particular that we can use π_{ABC} to quantify uncertainty in estimates" (F&P, p.5)

Calibrating ABC

> *"If π_{ABC} is calibrated, then this means that probability statements that are derived from it are appropriate, and in particular that we can use π_{ABC} to quantify uncertainty in estimates" (F&P, p.5)*

Definition

For $0 < q < 1$ and subset $\mathcal{A}$, event $E_q(\mathcal{A})$ made of $\mathbf{s}_{\text{obs}}$ such that $\text{Pr}_{\text{ABC}}(\theta \in \mathcal{A} | \mathbf{s}_{\text{obs}}) = q$. Then ABC is calibrated if

$$\text{Pr}(\theta \in \mathcal{A} | E_q(\mathcal{A})) = q$$

- unclear meaning of conditioning on $E_q(\mathcal{A})$

Calibrated ABC

Theorem (F&P)

Noisy ABC, where

$$\mathbf{s}_{\text{obs}} = S(\mathbf{y}_{\text{obs}}) + h\epsilon\,, \quad \epsilon \sim K(\cdot)$$

is calibrated

[Wilkinson, 2008]

no condition on h!!

Calibrated ABC

Consequence: when $h = \infty$

Theorem (F&P)

The prior distribution is always calibrated

is this a relevant property then?

More about calibrated ABC

> *"Calibration is not universally accepted by Bayesians. It is even more questionable here as we care how statements we make relate to the real world, not to a mathematically defined posterior."* R. Wilkinson

- Same reluctance about the prior being calibrated
- Property depending on prior, likelihood, and summary
- Calibration is a frequentist property (almost a p-value!)
- More sensible to account for the simulator's imperfections than using noisy-ABC against a meaningless based measure

[Wilkinson, 2012]

Converging ABC

Theorem (F&P)

For noisy ABC, the expected noisy-ABC log-likelihood,

$$\mathbb{E}\left\{\log[p(\theta|\mathbf{s}_{\text{obs}})]\right\} = \int\int \log[p(\theta|S(\mathbf{y}_{\text{obs}})+\epsilon)]\pi(\mathbf{y}_{\text{obs}}|\theta_0)K(\epsilon)\mathrm{d}\mathbf{y}_{\text{obs}}\mathrm{d}\epsilon,$$

has its maximum at $\theta = \theta_0$.

True for any choice of summary statistic? even ancilary statistics?!

[Imposes at least identifiability...]

Relevant in asymptotia and not for the data

Converging ABC

Corollary

For noisy ABC, the ABC posterior converges onto a point mass on the true parameter value as $m \to \infty$.

For standard ABC, not always the case (unless h goes to zero).

Strength of regularity conditions (c1) and (c2) in Bernardo & Smith, 1994?

[out-of-reach constraints on likelihood and posterior]

Again, there must be conditions imposed upon summary statistics...

Loss motivated statistic

Under quadratic loss function,

Theorem (F&P)

(i) The minimal posterior error $\mathbb{E}[L(\theta,\hat{\theta})|\mathbf{y}_{\text{obs}}]$ occurs when $\hat{\theta} = \mathbb{E}(\theta|\mathbf{y}_{\text{obs}})$ (!)

(ii) When $h \to 0$, $\mathbb{E}_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}})$ converges to $\mathbb{E}(\theta|\mathbf{y}_{\text{obs}})$

(iii) If $S(\mathbf{y}_{\text{obs}}) = \mathbb{E}[\theta|\mathbf{y}_{\text{obs}}]$ then for $\hat{\theta} = \mathbb{E}_{\text{ABC}}[\theta|\mathbf{s}_{\text{obs}}]$

$$\mathbb{E}[L(\theta,\hat{\theta})|\mathbf{y}_{\text{obs}}] = \text{trace}(A\Sigma) + h^2 \int \mathbf{x}^T A \mathbf{x} K(\mathbf{x}) d\mathbf{x} + o(h^2).$$

measure-theoretic difficulties?

dependence of $\mathbf{s}_{\text{obs}}$ on h ~~makes me uncomfortable~~ inherent to noisy ABC

Relevant for choice of K?

Optimal summary statistic

> *"We take a different approach, and weaken the requirement for π_{ABC} to be a good approximation to $\pi(\theta|\mathbf{y}_{obs})$. We argue for π_{ABC} to be a good approximation solely in terms of the accuracy of certain estimates of the parameters." (F&P, p.5)*

From this result, F&P

- derive their choice of summary statistic,

$$S(\mathbf{y}) = \mathbb{E}(\theta|\mathbf{y})$$

[almost sufficient]

- suggest

$$h = O(N^{-1/(2+d)}) \quad \text{and} \quad h = O(N^{-1/(4+d)})$$

as optimal bandwidths for noisy and standard ABC.

Optimal summary statistic

> *"We take a different approach, and weaken the requirement for π_{ABC} to be a good approximation to $\pi(\theta|\mathbf{y}_{obs})$. We argue for π_{ABC} to be a good approximation solely in terms of the accuracy of certain estimates of the parameters." (F&P, p.5)*

From this result, F&P

- derive their choice of summary statistic,

$$S(\mathbf{y}) = \mathbb{E}(\theta|\mathbf{y})$$

[wow! $\mathbb{E}_{\text{~~ABC~~}}[\theta|S(\mathbf{y}_{\text{obs}})] = \mathbb{E}[\theta|\mathbf{y}_{\text{obs}}]$]

- suggest

$$h = O(N^{-1/(2+d)}) \quad \text{and} \quad h = O(N^{-1/(4+d)})$$

as optimal bandwidths for noisy and standard ABC.

Caveat

Since $\mathbb{E}(\theta|\mathbf{y}_{\text{obs}})$ is most usually unavailable, F&P suggest

(i) use a pilot run of ABC to determine a region of non-negligible posterior mass;

(ii) simulate sets of parameter values and data;

(iii) use the simulated sets of parameter values and data to estimate the summary statistic; and

(iv) run ABC with this choice of summary statistic.

Monte Carl

Monday, October 32, 2010

Freed from likelihood!!!

Likelihood has always been a central part of our lifes and now that we can do without it, many wonder how to fill the vacuum created by this liberation. Other mutter it is anything but a liberation, as we now report from Monte Carlo.

Ten years ago, geneticists came with a technique they called ABC. Thanks to this miraculous recipe, they were able to overcome the dependence of simulation methods on the computation of the likelihood function, hence to overcome one of the most important difficulties in computational genetics.

The diffusion of the ABC technology has however been met with lukewarm enthusiasm by other members of the community and the whole Monte Carlo country has not yet been converted to this new creed. They will however be hard-pressed when faced with problems they cannot solve otherwise. One dearly hopes this resistance to progress does not lead to a schism by likelihoodists, or worse to guerilla actions to sabotage ABC!

Ren
foll
imp

The
that
rela
the
beh
of a
expr
in li
its
beh
con
or v

It m
tota
thin
dec

Caveat

Since $\mathbb{E}(\theta|\mathbf{y}_{\text{obs}})$ is most usually unavailable, F&P suggest

(i) use a pilot run of ABC to determine a region of non-negligible posterior mass;

(ii) simulate sets of parameter values and data;

(iii) use the simulated sets of parameter values and data to estimate the summary statistic; and

(iv) run ABC with this choice of summary statistic.

where is the assessment of the first stage error?

Monte Carl

Monday, October 32, 2010

Freed from likelihood!!!

Likelihood has always been a central part of our lifes and now that we can do without it, many wonder how to fill the vacuum created by this liberation. Other mutter it is anything but a liberation, as we now report from Monte Carlo.

Ten years ago, geneticists came with a technique they called ABC. Thanks to this miraculous recipe, they were able to overcome the dependence of simulation methods on the computation of the likelihood function, hence to overcome one of the most important difficulties in computational genetics.

The diffusion of the ABC technology has however been met with lukewarm enthusiasm by other members of the community and the whole Monte Carlo country has not yet been converted to this new creed. They will however be hard-pressed when faced with problems they cannot solve otherwise. One dearly hopes this resistance to progress does not lead to a schism by likelihoodists, or worse to guerilla actions to sabotage ABC!

Ren
foll
imp

The
that
rela
the
beh
of a
expr
in li
its
beh
con
or v

It m
tota
thin
dec

[my]questions about semi-automatic ABC

- dependence on h and $S(\cdot)$ in the early stage
- reduction of Bayesian inference to point estimation
- approximation error in step (i) not accounted for
- not parameterisation invariant
- practice shows that proper approximation to genuine posterior distributions stems from using a (much) larger number of summary statistics than the dimension of the parameter
- the validity of the approximation to the optimal summary statistic depends on the quality of the pilot run
- important inferential issues like model choice are not covered by this approach.

[Robert, 2012]

[my]questions about semi-automatic ABC

- dependence on h and $S(\cdot)$ in the early stage
- reduction of Bayesian inference to point estimation
- approximation error in step (i) not accounted for
- not parameterisation invariant
- practice shows that proper approximation to genuine posterior distributions stems from using a (much) larger number of summary statistics than the dimension of the parameter
- the validity of the approximation to the optimal summary statistic depends on the quality of the pilot run
- important inferential issues like model choice are not covered by this approach.

[X, 2012 ▸ get on with it!]

More about semi-automatic ABC

▸ End of section

[derived from comments on Read Paper, Series B, 2012]

“The apparently arbitrary nature of the choice of summary statistics has always been perceived as the Achilles heel of ABC.” M. Beaumont

More about semi-automatic ABC

[▸ End of section derived from comments on Read Paper, Series B, 2012]

"The apparently arbitrary nature of the choice of summary statistics has always been perceived as the Achilles heel of ABC." M. Beaumont

- "Curse of dimensionality" linked with the increase of the dimension of the summary statistic
- Connection with principal component analysis

  [Itan et al., 2010]
- Connection with partial least squares

  [Wegman et al., 2009]
- Beaumont et al. (2002) postprocessed output is used as input by F&P to run a second ABC

Wood's alternative

Instead of a non-parametric kernel approximation to the likelihood

$$\frac{1}{R}\sum_{r} K_\epsilon\{\eta(\mathbf{y}_r) - \eta(\mathbf{y}^{\text{obs}})\}$$

Wood (2010) suggests a normal approximation

$$\eta(\mathbf{y}(\theta)) \sim \mathcal{N}_d(\mu_\theta, \Sigma_\theta)$$

whose parameters can be approximated based on the R simulations (for each value of θ).

Wood's alternative

Instead of a non-parametric kernel approximation to the likelihood

$$\frac{1}{R}\sum_{r} K_{\epsilon}\{\eta(\mathbf{y}_r) - \eta(\mathbf{y}^{\text{obs}})\}$$

Wood (2010) suggests a normal approximation

$$\eta(\mathbf{y}(\theta)) \sim \mathcal{N}_d(\mu_\theta, \Sigma_\theta)$$

whose parameters can be approximated based on the R simulations (for each value of θ).

- Parametric versus non-parametric rate [Uh?!]
- Automatic weighting of components of $\eta(\cdot)$ through Σ_θ
- Dependence on normality assumption (pseudo-likelihood?)

[Cornebise, Girolami & Kosmidis, 2012]

Reinterpretation and extensions

Reinterpretation of ABC output as joint simulation from

$$\bar{\pi}(x, y|\theta) = f(x|\theta)\bar{\pi}_{Y|X}(y|x)$$

where

$$\bar{\pi}_{Y|X}(y|x) = K_\epsilon(y - x)$$

Reinterpretation of ABC output as joint simulation from

$$\bar{\pi}(x, y|\theta) = f(x|\theta)\bar{\pi}_{Y|X}(y|x)$$

where

$$\bar{\pi}_{Y|X}(y|x) = K_\epsilon(y - x)$$

Reinterpretation of noisy ABC

if $\bar{y}|y^{\text{obs}} \sim \bar{\pi}_{Y|X}(\cdot|y^{\text{obs}})$, then marginally

$$\bar{y} \sim \bar{\pi}_{Y|\theta}(\cdot|\theta^0)$$

© Explain for the consistency of Bayesian inference based on $\bar{y}$ and $\bar{\pi}$
[Lee, Andrieu & Doucet, 2012]

ABC for Markov chains

Rewriting the posterior as

$$\pi(\theta)^{1-n}\pi(\theta|x_1)\prod\pi(\theta|x_{t-1},x_t)$$

where $\pi(\theta|x_{t-1},x_t) \propto f(x_t|x_{t-1},\theta)\pi(\theta)$

ABC for Markov chains

Rewriting the posterior as

$$\pi(\theta)^{1-n}\pi(\theta|x_1)\prod\pi(\theta|x_{t-1},x_t)$$

where $\pi(\theta|x_{t-1},x_t) \propto f(x_t|x_{t-1},\theta)\pi(\theta)$

- Allows for a stepwise ABC, replacing each $\pi(\theta|x_{t-1},x_t)$ by an ABC approximation
- Similarity with F&P's multiple sources of data (and also with ◂ Dean et al., 2011)

[White et al., 2010, 2012]

Back to sufficiency

Difference between *regular sufficiency*, equivalent to

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

for all θ's and all priors π, and

Back to sufficiency

Difference between *regular sufficiency*, equivalent to

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

for all θ's and all priors π, and
marginal sufficiency, stated as

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

for all θ's, the given prior π and a subvector $\mu(\theta)$

[Basu, 1977]

Back to sufficiency

Difference between *regular sufficiency*, equivalent to

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

for all θ's and all priors π, and
marginal sufficiency, stated as

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

for all θ's, the given prior π and a subvector $\mu(\theta)$

[Basu, 1977]

Relates to F & P's main result, but could event be reduced to conditional sufficiency

$$\pi(\mu(\theta)|\mathbf{y}^{\text{obs}}) = \pi(\mu(\theta)|\eta(\mathbf{y}^{\text{obs}}))$$

(if feasible at all...)

[Dawson, 2012]

Predictive performances

Instead of posterior means, other aspects of posterior to explore.
E.g., look at minimising loss of information

$$\int p(\theta, \mathbf{y}) \log \frac{p(\theta, \mathbf{y})}{p(\theta)p(\mathbf{y})} \mathrm{d}\theta \mathrm{d}\mathbf{y} - \int p(\theta, \eta(\mathbf{y})) \log \frac{p(\theta, \eta(\mathbf{y}))}{p(\theta)p(\eta(\mathbf{y}))} \mathrm{d}\theta \mathrm{d}\eta(\mathbf{y})$$

for selection of summary statistics.

[Filippi, Barnes, & Stumpf, 2012]

Auxiliary variables

Auxiliary variable method avoids computations of untractable constant in likelihood

$$f(\mathbf{y}|\theta) = \mathfrak{Z}_\theta \tilde{f}(\mathbf{y}|\theta)$$

Auxiliary variables

Auxiliary variable method avoids computations of untractable constant in likelihood

$$f(\mathbf{y}|\theta) = \mathfrak{Z}_\theta \tilde{f}(\mathbf{y}|\theta)$$

Introduce pseudo-data $\mathbf{z}$ with artificial target $g(\mathbf{z}|\theta, \mathbf{y})$
Generate $\theta' \sim K(\theta, \theta')$ and $\mathbf{z}' \sim f(\mathbf{z}|\theta')$
Accept with probability

$$\frac{\pi(\theta')f(\mathbf{y}|\theta')g(\mathbf{z}'|\theta', \mathbf{y})}{\pi(\theta)f(\mathbf{y}|\theta)g(\mathbf{z}|\theta, \mathbf{y})} \frac{K(\theta', \theta)f(\mathbf{z}|\theta)}{K(\theta, \theta')f(\mathbf{z}'|\theta')} \wedge 1$$

[Møller, Pettitt, Berthelsen, & Reeves, 2006]

Auxiliary variables

Auxiliary variable method avoids computations of untractable constant in likelihood

$$f(\mathbf{y}|\theta) = \mathfrak{Z}_\theta \tilde{f}(\mathbf{y}|\theta)$$

Introduce pseudo-data $\mathbf{z}$ with artificial target $g(\mathbf{z}|\theta, \mathbf{y})$
Generate $\theta' \sim K(\theta, \theta')$ and $\mathbf{z}' \sim f(\mathbf{z}|\theta')$
Accept with probability

$$\frac{\pi(\theta')\tilde{f}(\mathbf{y}|\theta')g(\mathbf{z}'|\theta', \mathbf{y})}{\pi(\theta)\tilde{f}(\mathbf{y}|\theta)g(\mathbf{z}|\theta, \mathbf{y})} \frac{K(\theta', \theta)\tilde{f}(\mathbf{z}|\theta)}{K(\theta, \theta')\tilde{f}(\mathbf{z}'|\theta')} \wedge 1$$

[Møller, Pettitt, Berthelsen, & Reeves, 2006]

Auxiliary variables

Auxiliary variable method avoids computations of untractable constant in likelihood

$$f(\mathbf{y}|\theta) = \mathfrak{Z}_\theta \tilde{f}(\mathbf{y}|\theta)$$

Introduce pseudo-data $\mathbf{z}$ with artificial target $g(\mathbf{z}|\theta, \mathbf{y})$
Generate $\theta' \sim K(\theta, \theta')$ and $\mathbf{z}' \sim f(\mathbf{z}|\theta')$

For Gibbs random fields, existence of a genuine sufficient statistic $\eta(\mathbf{y})$.

[Møller, Pettitt, Berthelsen, & Reeves, 2006]

Auxiliary variables and ABC

Special case of ABC when

- $g(\mathbf{z}|\theta, \mathbf{y}) = K_\epsilon(\eta(\mathbf{z}) - \eta(\mathbf{y}))$
- $\tilde{f}(\mathbf{y}|\theta')\tilde{f}(\mathbf{z}|\theta)/\tilde{f}(\mathbf{y}|\theta)\tilde{f}(\mathbf{z}'|\theta')$ replaced by one [or not?!]

Auxiliary variables and ABC

Special case of ABC when

- $g(\mathbf{z}|\theta,\mathbf{y}) = K_\epsilon(\eta(\mathbf{z}) - \eta(\mathbf{y}))$
- $\tilde{f}(\mathbf{y}|\theta')\tilde{f}(\mathbf{z}|\theta)/\tilde{f}(\mathbf{y}|\theta)\tilde{f}(\mathbf{z}'|\theta')$ replaced by one [or not?!]

Consequences

- likelihood-free (ABC) versus constant-free (AVM)
- in ABC, $K_\epsilon(\cdot)$ should be allowed to depend on θ
- for Gibbs random fields, the auxiliary approach should be prefered to ABC

[Møller, 2012]

ABC and BIC

Idea of applying BIC during the ◂ local regression:

- Run regular ABC
- Select summary statistics during local regression
- Recycle the prior simulation sample (reference table) with those summary statistics
- Rerun the corresponding local regression (low cost)

[Pudlo & Sedki, 2012]

ABC for model choice

1. simulation-based methods in Econometrics
2. Genetics of ABC
3. Approximate Bayesian computation
4. ABC for model choice
5. ABC model choice via random forests
6. ABC estimation via random forests
7. [some] asymptotics of ABC

Bayesian model choice

Several models $M_1, M_2, \ldots$ are considered simultaneously for a dataset $\mathbf{y}$ and the model index $\mathcal{M}$ is part of the inference.
Use of a prior distribution. $\pi(\mathcal{M} = m)$, plus a prior distribution on the parameter conditional on the value m of the model index, $\pi_m(\boldsymbol{\theta}_m)$

Goal is to derive the posterior distribution of M, challenging computational target when models are complex.

Generic ABC for model choice

Algorithm 4 Likelihood-free model choice sampler (ABC-MC)

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

ABC estimates

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

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

Issues with implementation:

- should tolerances ϵ be the same for all models?
- should summary statistics vary across models (incl. their dimension)?
- should the distance measure ρ vary as well?

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]

The Great ABC controversy

On-going controvery in phylogeographic genetics about the validity of using ABC for testing

Against: Templeton, 2008, 2009, 2010a, 2010b, 2010c argues that nested hypotheses cannot have higher probabilities than nesting hypotheses (!)

fodey.com

The Great ABC controversy

On-going controvery in phylogeographic genetics about the validity of using ABC for testing

Against: Templeton, 2008, 2009, 2010a, 2010b, 2010c argues that nested hypotheses cannot have higher probabilities than nesting hypotheses (!)

Replies: Fagundes et al., 2008, Beaumont et al., 2010, Berger et al., 2010, Csillèry et al., 2010 point out that the criticisms are addressed at [Bayesian] model-based inference and have nothing to do with ABC...

Gibbs random fields

Gibbs distribution

The rv $\mathbf{y} = (y_1, \ldots, y_n)$ is a **Gibbs random field** associated with the graph $\mathfrak{G}$ if

$$f(\mathbf{y}) = \frac{1}{\mathfrak{Z}} \exp\left\{ -\sum_{c \in \mathscr{C}} V_c(\mathbf{y}_c) \right\} ,$$

where $\mathfrak{Z}$ is the normalising constant, $\mathscr{C}$ is the set of cliques of $\mathfrak{G}$ and V_c is any function also called **potential** ◂ sufficient statistic
$U(\mathbf{y}) = \sum_{c \in \mathscr{C}} V_c(\mathbf{y}_c)$ is the **energy function**

Gibbs random fields

Gibbs distribution

The rv $\mathbf{y} = (y_1, \ldots, y_n)$ is a **Gibbs random field** associated with the graph $\mathfrak{G}$ if

$$f(\mathbf{y}) = \frac{1}{\mathfrak{Z}} \exp\left\{ - \sum_{c \in \mathscr{C}} V_c(\mathbf{y}_c) \right\} ,$$

where $\mathfrak{Z}$ is the normalising constant, $\mathscr{C}$ is the set of cliques of $\mathfrak{G}$ and V_c is any function also called **potential** ◂ sufficient statistic
$U(\mathbf{y}) = \sum_{c \in \mathscr{C}} V_c(\mathbf{y}_c)$ is the **energy function**

ⓒ $\mathfrak{Z}$ is usually unavailable in closed form

Potts model

$V_c(\mathbf{y})$ is of the form

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

where $l \sim i$ denotes a neighbourhood structure

Potts model

Potts model

$V_c(\mathbf{y})$ is of the form

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

where $l \sim i$ denotes a neighbourhood structure

In most realistic settings, summation

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

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

[Cucala, Marin, CPR & Titterington, 2009]

Comparing a model with potential S_0 taking values in $\mathbb{R}^{p_0}$ versus a model with potential S_1 taking values in $\mathbb{R}^{p_1}$ can be done through the Bayes factor corresponding to the priors π_0 and π_1 on each parameter space

$$\mathfrak{B}_{m_0/m_1}(\mathbf{x}) = \frac{\int \exp\{\boldsymbol{\theta}_0^{\mathsf{T}} S_0(\mathbf{x})\}/Z_{\boldsymbol{\theta}_0,0}\pi_0(\mathrm{d}\boldsymbol{\theta}_0)}{\int \exp\{\boldsymbol{\theta}_1^{\mathsf{T}} S_1(\mathbf{x})\}/Z_{\boldsymbol{\theta}_1,1}\pi_1(\mathrm{d}\boldsymbol{\theta}_1)}$$

Bayesian Model Choice

Comparing a model with potential S_0 taking values in $\mathbb{R}^{p_0}$ versus a model with potential S_1 taking values in $\mathbb{R}^{p_1}$ can be done through the Bayes factor corresponding to the priors π_0 and π_1 on each parameter space

$$\mathfrak{B}_{m_0/m_1}(\mathbf{x}) = \frac{\int \exp\{\boldsymbol{\theta}_0^{\mathsf{T}} S_0(\mathbf{x})\}/Z_{\boldsymbol{\theta}_0,0}\pi_0(\mathrm{d}\boldsymbol{\theta}_0)}{\int \exp\{\boldsymbol{\theta}_1^{\mathsf{T}} S_1(\mathbf{x})\}/Z_{\boldsymbol{\theta}_1,1}\pi_1(\mathrm{d}\boldsymbol{\theta}_1)}$$

Use of Jeffreys' scale to select most appropriate model

Neighbourhood relations

Choice to be made between M neighbourhood relations

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

with

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

driven by the posterior probabilities of the models.

Model index

Formalisation via a model index $\mathcal{M}$ that appears as a new parameter with prior distribution $\pi(\mathcal{M} = m)$ and $\pi(\theta|\mathcal{M} = m) = \pi_m(\theta_m)$

Model index

Formalisation via a model index $\mathcal{M}$ that appears as a new parameter with prior distribution $\pi(\mathcal{M}=m)$ and $\pi(\theta|\mathcal{M}=m)=\pi_m(\theta_m)$
Computational target:

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

Sufficient statistics

By definition, 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

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

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

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

Sufficient statistics in Gibbs random fields

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

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

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

[Specific to Gibbs random fields!]

ABC model choice Algorithm

ABC-MC

- Generate m^* from the prior $\pi(\mathcal{M} = m)$.
- Generate $\theta^*_{m^*}$ from the prior $\pi_{m^*}(\cdot)$.
- Generate x^* from the model $f_{m^*}(\cdot|\theta^*_{m^*})$.
- Compute the distance $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*))$.
- Accept $(\theta^*_{m^*}, m^*)$ if $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon$.

Note When $\epsilon = 0$ the algorithm is exact

ABC approximation to the Bayes factor

Frequency ratio:

$$\begin{aligned}
\overline{BF}_{m_0/m_1}(\mathbf{x}^0) &= \frac{\hat{\mathbb{P}}(\mathcal{M}=m_0|\mathbf{x}^0)}{\hat{\mathbb{P}}(\mathcal{M}=m_1|\mathbf{x}^0)} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)} \\
&= \frac{\sharp\{m^{i*}=m_0\}}{\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}\,,
\end{aligned}$$

ABC approximation to the Bayes factor

Frequency ratio:

$$\begin{aligned}\overline{BF}_{m_0/m_1}(\mathbf{x}^0) &= \frac{\hat{\mathbb{P}}(\mathcal{M}=m_0|\mathbf{x}^0)}{\hat{\mathbb{P}}(\mathcal{M}=m_1|\mathbf{x}^0)} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)} \\ &= \frac{\sharp\{m^{i*}=m_0\}}{\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}\,,\end{aligned}$$

replaced with

$$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{1+\sharp\{m^{i*}=m_0\}}{1+\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}$$

to avoid indeterminacy (also Bayes estimate).

Toy example

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

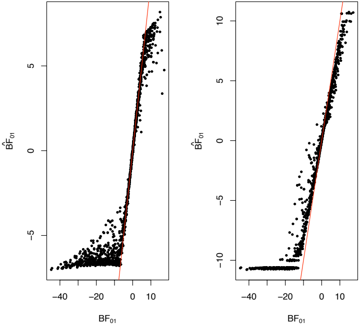
$$f_0(\mathbf{x}|\theta_0) = \exp\left(\theta_0 \sum_{i=1}^{n} \mathbb{I}_{\{x_i=1\}}\right) \Big/ \{1+\exp(\theta_0)\}^n ,$$

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).

Toy example (2)



(left) Comparison of the true $BF_{m_0/m_1}(\mathbf{x}^0)$ with $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ (in logs) over $2,000$ simulations and 4.10^6 proposals from the prior. *(right)* Same when using tolerance ϵ corresponding to the 1% quantile on the distances.

Back to sufficiency

> *‘Sufficient statistics for individual models are unlikely to be very informative for the model probability.’*
>
> [Scott Sisson, Jan. 31, 2011, X.’Og]

Back to sufficiency

> *'Sufficient statistics for individual models are unlikely to be very informative for the model probability.'*
>
> [Scott Sisson, Jan. 31, 2011, X.'Og]

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

Back to sufficiency

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

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

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

© Potential loss of information at the testing level

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

ABC approximation

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

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

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

ABC approximation

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

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

$$\begin{aligned} B_{12}^{\epsilon}(\mathbf{y}) &= \frac{\int \mathbb{I}_{\rho\{\eta(\mathbf{z}),\eta(\mathbf{y})\}\le\epsilon}\pi_1(\boldsymbol{\theta}_1) f_1(\mathbf{z}|\boldsymbol{\theta}_1)\,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\theta}_1}{\int \mathbb{I}_{\rho\{\eta(\mathbf{z}),\eta(\mathbf{y})\}\le\epsilon}\pi_2(\boldsymbol{\theta}_2) f_2(\mathbf{z}|\boldsymbol{\theta}_2)\,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\theta}_2} \\ &= \frac{\int \mathbb{I}_{\rho\{\eta,\eta(\mathbf{y})\}\le\epsilon}\pi_1(\boldsymbol{\theta}_1) f_1^{\eta}(\eta|\boldsymbol{\theta}_1)\,\mathrm{d}\eta\,\mathrm{d}\boldsymbol{\theta}_1}{\int \mathbb{I}_{\rho\{\eta,\eta(\mathbf{y})\}\le\epsilon}\pi_2(\boldsymbol{\theta}_2) f_2^{\eta}(\eta|\boldsymbol{\theta}_2)\,\mathrm{d}\eta\,\mathrm{d}\boldsymbol{\theta}_2}\,, \end{aligned}$$

where $f_1^{\eta}(\eta|\boldsymbol{\theta}_1)$ and $f_2^{\eta}(\eta|\boldsymbol{\theta}_2)$ distributions of $\eta(\mathbf{z})$

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

When ϵ goes to zero,

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

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

When ϵ goes to zero,

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

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

Limiting behaviour of B_{12} (under sufficiency)

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

$$f_i(\mathbf{y}|\boldsymbol{\theta}_i) = g_i(\mathbf{y}) f_i^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_i)$$

Thus

$$\begin{aligned}
B_{12}(\mathbf{y}) &= \frac{\int_{\Theta_1} \pi(\boldsymbol{\theta}_1) g_1(\mathbf{y}) f_1^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1}{\int_{\Theta_2} \pi(\boldsymbol{\theta}_2) g_2(\mathbf{y}) f_2^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2} \\
&= \frac{g_1(\mathbf{y})\ \int \pi_1(\boldsymbol{\theta}_1) f_1^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1}{g_2(\mathbf{y})\ \int \pi_2(\boldsymbol{\theta}_2) f_2^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2} = \frac{g_1(\mathbf{y})}{g_2(\mathbf{y})}\, B_{12}^\eta(\mathbf{y})\,.
\end{aligned}$$

[Didelot, Everitt, Johansen & Lawson, 2011]

Limiting behaviour of B_{12} (under sufficiency)

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

$$f_i(\mathbf{y}|\boldsymbol{\theta}_i) = g_i(\mathbf{y}) f_i^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_i)$$

Thus

$$\begin{aligned} B_{12}(\mathbf{y}) &= \frac{\int_{\Theta_1} \pi(\boldsymbol{\theta}_1) g_1(\mathbf{y}) f_1^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1}{\int_{\Theta_2} \pi(\boldsymbol{\theta}_2) g_2(\mathbf{y}) f_2^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2} \\ &= \frac{g_1(\mathbf{y})\ \int \pi_1(\boldsymbol{\theta}_1) f_1^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1}{g_2(\mathbf{y})\ \int \pi_2(\boldsymbol{\theta}_2) f_2^\eta(\eta(\mathbf{y})|\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2} = \frac{g_1(\mathbf{y})}{g_2(\mathbf{y})}\, B_{12}^\eta(\mathbf{y})\,. \end{aligned}$$

[Didelot, Everitt, Johansen & Lawson, 2011]

© No discrepancy only when cross-model sufficiency

Poisson/geometric example

Sample

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

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

$$S = \sum_{i=1}^{n} y_i = \eta(\mathbf{x})$$

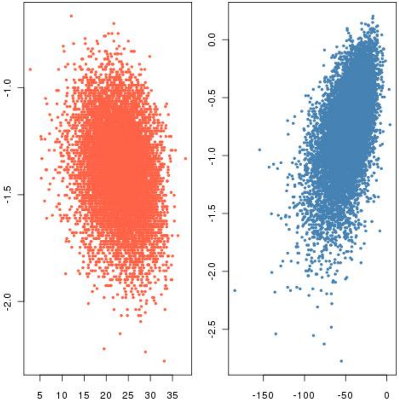
sufficient statistic for either model but not simultaneously

Discrepancy ratio

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

Poisson/geometric discrepancy

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



Formal recovery

Creating an encompassing exponential family

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

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]

Formal recovery

Creating an encompassing exponential family

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

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

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

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 loss brought by summary statistics**

Meaning of the ABC-Bayes factor

> *'This is also why focus on model discrimination typically (...) proceeds by (...) accepting that the Bayes Factor that one obtains is only derived from the summary statistics and may in no way correspond to that of the full model.'*
>
> [Scott Sisson, Jan. 31, 2011, X.'Og]

Meaning of the ABC-Bayes factor

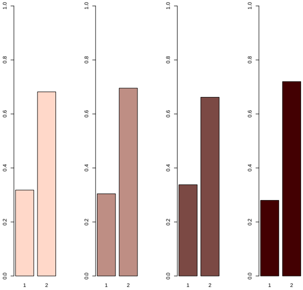
> *'This is also why focus on model discrimination typically (...) proceeds by (...) accepting that the Bayes Factor that one obtains is only derived from the summary statistics and may in no way correspond to that of the full model.'*

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

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

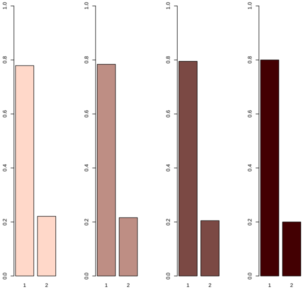
$$\lim_{n\to\infty} B_{12}^{\eta}(\mathbf{y}) = \frac{(\theta_0 + 1)^2}{\theta_0} e^{-\theta_0}$$

MA(q) divergence



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(q) divergence



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.

Further comments

> *'There should be the possibility that for the same model, but different (non-minimal) [summary] statistics (so different η's: η_1 and η_1^*) the ratio of evidences may no longer be equal to one.'*

[Michael Stumpf, Jan. 28, 2011, 'Og]

Using different summary statistics [on different models] may indicate the loss of information brought by each set but agreement does not lead to trustworthy approximations.

A stylised problem

Central question to the validation of ABC for model choice:

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

A stylised problem

Central question to the validation of ABC for model choice:

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

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

[Marin, Pillai, X, & Rousseau, JRSS B, 2013]

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

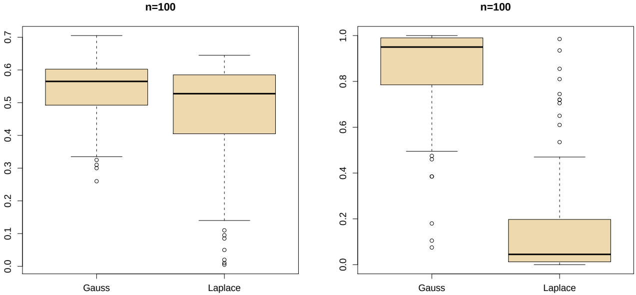
1. sample mean $\bar{\mathbf{y}}$ (sufficient for $\mathfrak{M}_1$ if not $\mathfrak{M}_2$);
2. sample median $\mathrm{med}(\mathbf{y})$ (insufficient);
3. sample variance $\mathrm{var}(\mathbf{y})$ (ancillary);
4. median absolute deviation $\mathrm{mad}(\mathbf{y}) = \mathrm{med}(|\mathbf{y} - \mathrm{med}(\mathbf{y})|)$;

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).



Framework

Starting from sample

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

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

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

Summary statistics

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

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

Framework

ⓒ Comparison of

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

turned into

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

Assumptions

A collection of asymptotic “standard” assumptions:

[A1] is a standard central limit theorem under the true model with asymptotic mean μ_0
[A2] controls the large deviations of the estimator $\boldsymbol{T}^n$ from the model mean $\mu(\theta)$
[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!]

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

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

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

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

Between-model consistency

Consequence of above is that asymptotic behaviour of the Bayes factor is driven by the asymptotic mean value $\mu(\theta)$ of $\boldsymbol{T}^n$ 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 $\mu(\theta)$ of $\boldsymbol{T}^n$ 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

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

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 $\mu(\theta)$ of $\boldsymbol{T}^n$ 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

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

Checking for adequate statistics

Run a practical check of the relevance (or non-relevance) of $\boldsymbol{T}^n$

null hypothesis that both models are compatible with the statistic $\boldsymbol{T}^n$

$$H_0 : \inf\{|\mu_2(\theta_2) - \mu_0|; \theta_2 \in \Theta_2\} = 0$$

against

$$H_1 : \inf\{|\mu_2(\theta_2) - \mu_0|; \theta_2 \in \Theta_2\} > 0$$

testing procedure provides estimates of mean of $\boldsymbol{T}^n$ under each model and checks for equality

Checking in practice

- Under each model $\mathfrak{M}_i$, generate ABC sample $\theta_{i,l}, l = 1, \cdots, L$
- For each $\theta_{i,l}$, generate $\mathbf{y}_{i,l} \sim F_{i,n}(\cdot|\psi_{i,l})$, derive $\boldsymbol{T}^n(\mathbf{y}_{i,l})$ and compute

$$\hat{\mu}_i = \frac{1}{L} \sum_{l=1}^{L} \boldsymbol{T}^n(\mathbf{y}_{i,l}), \quad i = 1, 2\,.$$

- Conditionally on $\boldsymbol{T}^n(\mathbf{y})$,

$$\sqrt{L}\,\{\hat{\mu}_i - \mathbb{E}^\pi\left[\mu_i(\theta_i)|\,\boldsymbol{T}^n(\mathbf{y})\right]\} \rightsquigarrow \mathcal{N}(0, V_i),$$

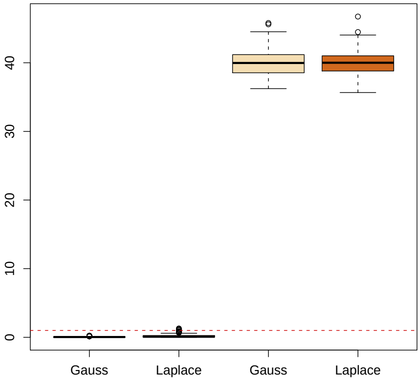
- Test for a common mean

$$H_0 : \hat{\mu}_1 \sim \mathcal{N}(\mu_0, V_1)\,, \hat{\mu}_2 \sim \mathcal{N}(\mu_0, V_2)$$

against the alternative of different means

$$H_1 : \hat{\mu}_i \sim \mathcal{N}(\mu_i, V_i), \quad \text{with } \mu_1 \neq \mu_2\,.$$

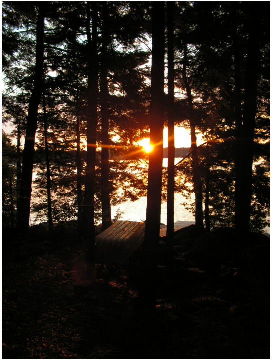
Toy example: Laplace versus Gauss



Normalised χ^2 without and with mad

ABC model choice via random forests

1. simulation-based methods in Econometrics
2. Genetics of ABC
3. Approximate Bayesian computation
4. ABC for model choice
5. ABC model choice via random forests
   - Random forests
   - ABC with random forests
   - Illustrations
6. ABC estimation via random forests
7. [some] asymptotics of ABC



Leaning towards machine learning

Main notions:

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

[Cornuet et al., 2014, in progress]

Random forests

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

[Breiman, 1996]

Improved classification performances by averaging over classification schemes of randomly generated training sets, creating a "forest" of (CART) decision trees, inspired by Amit and Geman (1997) ensemble learning

[Breiman, 2001]

Growing the forest

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

- boostrap resampling of the dataset and
- random subset-ing [of size $\sqrt{t}$] of the covariates driving the classification at every node of each tree

Covariate x_τ that drives the node separation

$$x_\tau \gtrless c_\tau$$

and the separation bound c_τ chosen by minimising entropy or Gini index

Breiman and Cutler's algorithm

Algorithm 5 Random forests

```
for $t = 1$ to $T$ do
    //*T is the number of trees*//
    Draw a bootstrap sample of size $n_{\text{boot}}$
    Grow an unpruned decision tree
    for $b = 1$ to $B$ do
        //*B is the number of nodes*//
        Select $n_{\text{try}}$ of the predictors at random
        Determine the best split from among those predictors
    end for
end for
Predict new data by aggregating the predictions of the $T$ trees
```

Subsampling

Due to both large datasets [practical] and theoretical recommendation from Gérard Biau [private communication], from independence between trees to convergence issues, boostrap sample of much smaller size than original data size

$$N = \mathrm{o}(n)$$

Subsampling

Due to both large datasets [practical] and theoretical recommendation from Gérard Biau [private communication], from independence between trees to convergence issues, boostrap sample of much smaller size than original data size

$$N = \mathrm{o}(n)$$

Each CART tree stops when number of observations per node is 1: no culling of the branches

ABC with random forests

Idea: Starting with

- possibly large collection of summary statistics $(s_{1i}, \ldots, s_{pi})$ (from scientific theory input to available statistical softwares, to machine-learning alternatives)
- ABC reference table involving model index, parameter values and summary statistics for the associated simulated pseudo-data

run R randomforest to infer $\mathfrak{M}$ from $(s_{1i}, \ldots, s_{pi})$

ABC with random forests

Idea: Starting with

- possibly large collection of summary statistics $(s_{1i}, \ldots, s_{pi})$ (from scientific theory input to available statistical softwares, to machine-learning alternatives)
- ABC reference table involving model index, parameter values and summary statistics for the associated simulated pseudo-data

run R randomforest to infer $\mathfrak{M}$ from $(s_{1i}, \ldots, s_{pi})$

at each step $\mathrm{O}(\sqrt{p})$ indices sampled at random and most discriminating statistic selected, by minimising ~~entropy~~ Gini loss

ABC with random forests

Idea: Starting with

- possibly large collection of summary statistics $(s_{1i}, \ldots, s_{pi})$ (from scientific theory input to available statistical softwares, to machine-learning alternatives)
- ABC reference table involving model index, parameter values and summary statistics for the associated simulated pseudo-data

run R randomforest to infer $\mathfrak{M}$ from $(s_{1i}, \ldots, s_{pi})$

Average of the trees is resulting summary statistics, highly non-linear predictor of the model index

Outcome of ABC-RF

Random forest predicts a (MAP) model index, from the observed dataset: The predictor provided by the forest is “sufficient” to select the most likely model but not to derive associated posterior probability

Outcome of ABC-RF

Random forest predicts a (MAP) model index, from the observed dataset: The predictor provided by the forest is “sufficient” to select the most likely model but not to derive associated posterior probability

- exploit entire forest by computing how many trees lead to picking each of the models under comparison but variability too high to be trusted
- frequency of trees associated with majority model is no proper substitute to the true posterior probability
- And usual ABC-MC approximation equally highly variable and hard to assess

Posterior predictive expected losses

We suggest replacing unstable approximation of

$$\mathbb{P}(\mathfrak{M} = m | x_o)$$

with x_o observed sample and m model index, by average of the selection errors across all models given the data x_o,

$$\mathbb{P}(\hat{\mathfrak{M}}(X) \neq \mathfrak{M} | x_o)$$

where pair $(\mathfrak{M}, X)$ generated from the predictive

$$\int f(x|\theta) \pi(\theta, \mathfrak{M} | x_o) \mathrm{d}\theta$$

and $\hat{\mathfrak{M}}(x)$ denotes the random forest model (MAP) predictor

Posterior predictive expected losses

Arguments:

- Bayesian estimate of the posterior error
- integrates error over most likely part of the parameter space
- gives an averaged error rather than the posterior probability of the null hypothesis
- easily computed: Given ABC subsample of parameters from reference table, simulate pseudo-samples associated with those and derive error frequency

toy: MA(1) vs. MA(2)

Comparing an MA(1) and an MA(2) models:

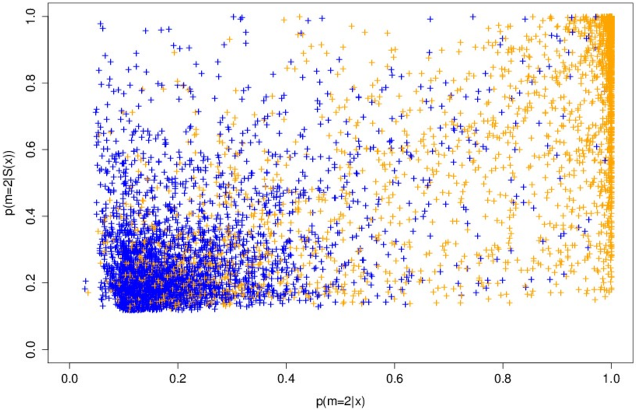
$$x_t = \epsilon_t - \vartheta_1 \epsilon_{t-1} [-\vartheta_2 \epsilon_{t-2}]$$

Earlier illustration using first two autocorrelations as $S(x)$

[Marin et al., Stat. & Comp., 2011]

Result #1: values of $p(m|x)$ [obtained by numerical integration] and $p(m|S(x))$ [obtained by mixing ABC outcome and density estimation] highly differ!

toy: MA(1) vs. MA(2)



Difference between the posterior probability of $MA(2)$ given either x or $S(x)$. Blue stands for data from $MA(1)$, orange for data from $MA(2)$

toy: MA(1) vs. MA(2)

Comparing an MA(1) and an MA(2) models:

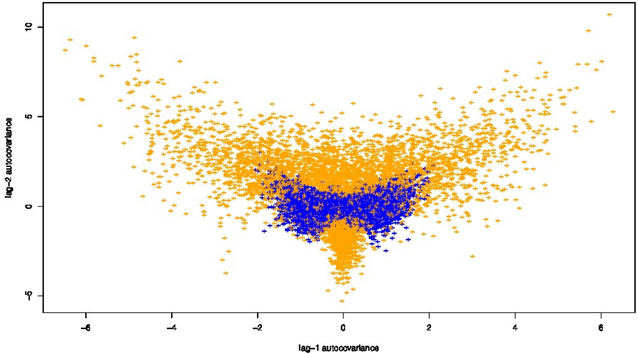
$$x_t = \epsilon_t - \vartheta_1 \epsilon_{t-1} [-\vartheta_2 \epsilon_{t-2}]$$

Earlier illustration using two autocorrelations as $S(x)$

[Marin et al., Stat. & Comp., 2011]

Result #2: Embedded models, with simulations from MA(1) within those from MA(2), hence linear classification poor

toy: MA(1) vs. MA(2)



Simulations of $S(x)$ under $MA(1)$ (blue) and $MA(2)$ (orange)

toy: MA(1) vs. MA(2)

Comparing an MA(1) and an MA(2) models:

$$x_t = \epsilon_t - \vartheta_1 \epsilon_{t-1} [-\vartheta_2 \epsilon_{t-2}]$$

Earlier illustration using two autocorrelations as $S(x)$

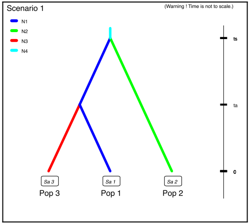
[Marin et al., Stat. & Comp., 2011]

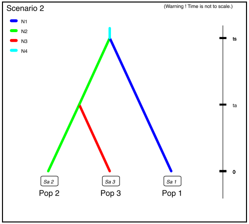
Result #3: On such a small dimension problem, random forests should come second to *k*-nn ou kernel discriminant analyses

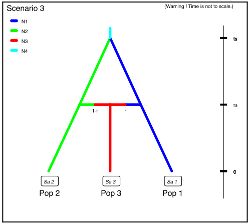
toy: MA(1) vs. MA(2)

classification **method**	**prior** **error rate** (in %)
LDA	27.43
Logist. reg.	28.34
SVM (library e1071)	17.17
“naïve” Bayes (with G marg.)	19.52
“naïve” Bayes (with NP marg.)	18.25
ABC k-nn ($k = 100$)	17.23
ABC k-nn ($k = 50$)	16.97
Local log. reg. ($k = 1000$)	16.82
Random Forest	17.04
Kernel disc. ana. (KDA)	16.95
True MAP	12.36

Evolution scenarios based on SNPs







Three scenarios for the evolution of three populations from their most common ancestor

DIYBAC header (!)

```
7 parameters and 48 summary statistics

3 scenarios: 7 7 7
scenario 1 [0.33333] (6)
N1 N2 N3
0 sample 1
0 sample 2
0 sample 3
ta merge 1 3
ts merge 1 2
ts varne 1 N4
scenario 2 [0.33333] (6)
N1 N2 N3
..........
ts varne 1 N4
scenario 3 [0.33333] (7)
N1 N2 N3
........
historical parameters priors (7,1)
N1 N UN[100.0,30000.0,0.0,0.0]
N2 N UN[100.0,30000.0,0.0,0.0]
N3 N UN[100.0,30000.0,0.0,0.0]
ta T UN[10.0,30000.0,0.0,0.0]
ts T UN[10.0,30000.0,0.0,0.0]
N4 N UN[100.0,30000.0,0.0,0.0]
r A UN[0.05,0.95,0.0,0.0]
ts>ta
DRAW UNTIL
```

Evolution scenarios based on SNPs

Model 1 with 6 parameters:

- four effective sample sizes: N_1 for population 1, N_2 for population 2, N_3 for population 3 and, finally, N_4 for the native population;
- the time of divergence t_a between populations 1 and 3;
- the time of divergence t_s between populations 1 and 2.
- effective sample sizes with independent uniform priors on $[100, 30000]$
- vector of divergence times (t_a, t_s) with uniform prior on $\{(a, s) \in [10, 30000] \otimes [10, 30000] | a < s\}$

Evolution scenarios based on SNPs

Model 2 with same parameters as model 1 but the divergence time t_a corresponds to a divergence between populations 2 and 3; prior distributions identical to those of model 1
Model 3 with extra seventh parameter, admixture rate r. For that scenario, at time t_a admixture between populations 1 and 2 from which population 3 emerges. Prior distribution on r uniform on $[0.05, 0.95]$. In that case models 1 and 2 are not embeddeded in model 3. Prior distributions for other parameters the same as in model 1

Evolution scenarios based on SNPs

Set of 48 summary statistics:

Single sample statistics

- proportion of loci with null gene diversity (= proportion of monomorphic loci)
- mean gene diversity across polymorphic loci

[Nei, 1987]

- variance of gene diversity across polymorphic loci
- mean gene diversity across all loci

Evolution scenarios based on SNPs

Set of 48 summary statistics:

Two sample statistics

- proportion of loci with null FST distance between both samples [Weir and Cockerham, 1984]
- mean across loci of non null FST distances between both samples
- variance across loci of non null FST distances between both samples
- mean across loci of FST distances between both samples
- proportion of 1 loci with null Nei's distance between both samples [Nei, 1972]
- mean across loci of non null Nei's distances between both samples
- variance across loci of non null Nei's distances between both samples
- mean across loci of Nei's distances between the two samples

Evolution scenarios based on SNPs

Set of 48 summary statistics:

Three sample statistics

- proportion of loci with null admixture estimate
- mean across loci of non null admixture estimate
- variance across loci of non null admixture estimated
- mean across all locus admixture estimates

Evolution scenarios based on SNPs

For a sample of 1000 SNIPs measured on 25 biallelic individuals per population, learning ABC reference table with $20,000$ simulations, prior predictive error rates:

- “naïve Bayes” classifier 33.3%
- raw LDA classifier 23.27%
- ABC k-nn [Euclidean dist. on summaries normalised by MAD] 25.93%
- ABC k-nn [unnormalised Euclidean dist. on LDA components] 22.12%
- local logistic classifier based on LDA components with
  - $k = 500$ neighbours 22.61%
- random forest on summaries 21.03%

(Error rates computed on a prior sample of size 10^4)

Evolution scenarios based on SNPs

For a sample of 1000 SNIPs measured on 25 biallelic individuals per population, learning ABC reference table with $20,000$ simulations, prior predictive error rates:

- "naïve Bayes" classifier 33.3%
- raw LDA classifier 23.27%
- ABC k-nn [Euclidean dist. on summaries normalised by MAD] 25.93%
- ABC k-nn [unnormalised Euclidean dist. on LDA components] 22.12%
- local logistic classifier based on LDA components with
    - $k = 1000$ neighbours 22.46%
- random forest on summaries 21.03%

(Error rates computed on a prior sample of size 10^4)

Evolution scenarios based on SNPs

For a sample of 1000 SNIPs measured on 25 biallelic individuals per population, learning ABC reference table with $20,000$ simulations, prior predictive error rates:

- "naïve Bayes" classifier 33.3%
- raw LDA classifier 23.27%
- ABC k-nn [Euclidean dist. on summaries normalised by MAD] 25.93%
- ABC k-nn [unnormalised Euclidean dist. on LDA components] 22.12%
- local logistic classifier based on LDA components with
  - $k = 5000$ neighbours 22.43%
- random forest on summaries 21.03%

(Error rates computed on a prior sample of size 10^4)

Evolution scenarios based on SNPs

For a sample of 1000 SNIPs measured on 25 biallelic individuals per population, learning ABC reference table with $20,000$ simulations, prior predictive error rates:

- "naïve Bayes" classifier   33.3%
- raw LDA classifier   23.27%
- ABC k-nn [Euclidean dist. on summaries normalised by MAD] 25.93%
- ABC k-nn [unnormalised Euclidean dist. on LDA components] 22.12%
- local logistic classifier based on LDA components with
  - $k = 5000$ neighbours   22.43%
- random forest on LDA components only   23.1%

(Error rates computed on a prior sample of size 10^4)

Evolution scenarios based on SNPs

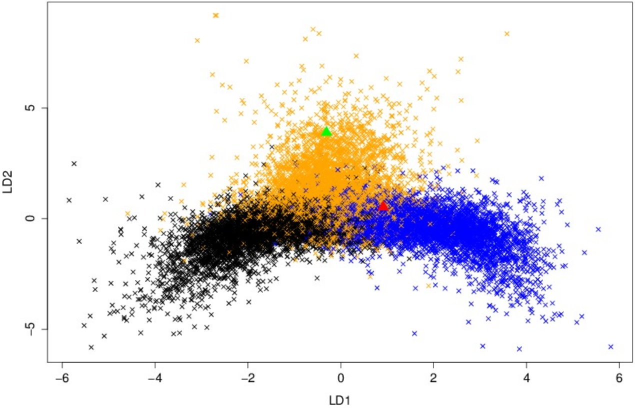
For a sample of 1000 SNIPs measured on 25 biallelic individuals per population, learning ABC reference table with $20,000$ simulations, prior predictive error rates:

- "naïve Bayes" classifier 33.3%
- raw LDA classifier 23.27%
- ABC k-nn [Euclidean dist. on summaries normalised by MAD] 25.93%
- ABC k-nn [unnormalised Euclidean dist. on LDA components] 22.12%
- local logistic classifier based on LDA components with
  - $k = 5000$ neighbours 22.43%
- random forest on summaries and LDA components 19.03%

(Error rates computed on a prior sample of size 10^4)

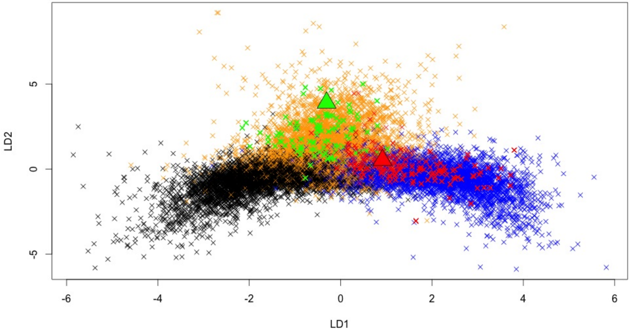
Evolution scenarios based on SNPs

Posterior predictive error rates



Evolution scenarios based on SNPs

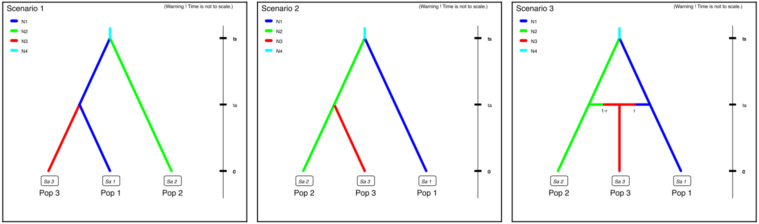
Posterior predictive error rates



favourable: 0.010 error – unfavourable: 0.104 error

Evolution scenarios based on microsatellites

Same setting as previously



Sample of 25 diploid individuals per population, on 20 locus (roughly corresponds to 1/5th of previous information)

Evolution scenarios based on microsatellites

One sample statistics

- mean number of alleles across loci
- mean gene diversity across loci (Nei, 1987)
- mean allele size variance across loci
- mean M index across loci (Garza and Williamson, 2001; Excoffier et al., 2005)

Evolution scenarios based on microsatellites

Two sample statistics

- mean number of alleles across loci (two samples)
- mean gene diversity across loci (two samples)
- mean allele size variance across loci (two samples)
- FST between two samples (Weir and Cockerham, 1984)
- mean index of classification (two samples) (Rannala and Moutain, 1997; Pascual et al., 2007)
- shared allele distance between two samples (Chakraborty and Jin, 1993)
- $(\delta\mu)^2$ distance between two samples (Golstein et al., 1995)

Three sample statistics

- Maximum likelihood coefficient of admixture (Choisy et al., 2004)

Evolution scenarios based on microsatellites

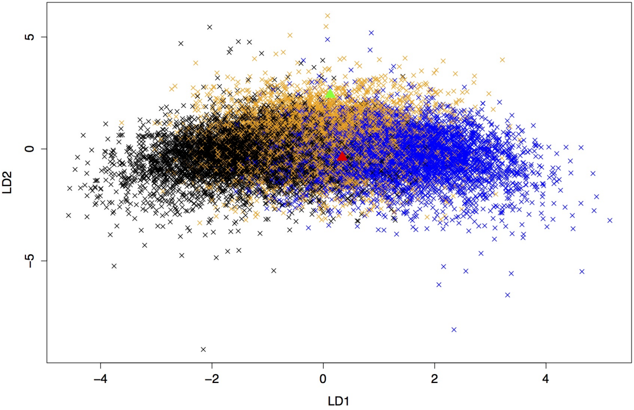
classification method	**prior error**[*] **rate** (in %)
raw LDA	35.64
"naïve" Bayes (with G marginals)	40.02
k-nn (MAD normalised sum stat)	37.47
k-nn (unormalised LDA)	35.14
RF without LDA components	35.14
RF with LDA components	33.62
RF with only LDA components	37.25

---

[*] estimated on pseudo-samples of 10^4 items drawn from the prior

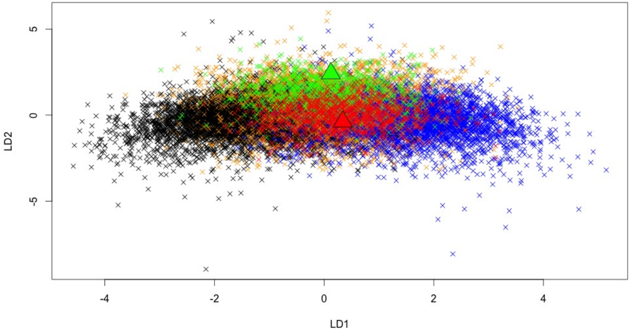
Evolution scenarios based on microsatellites

Posterior predictive error rates



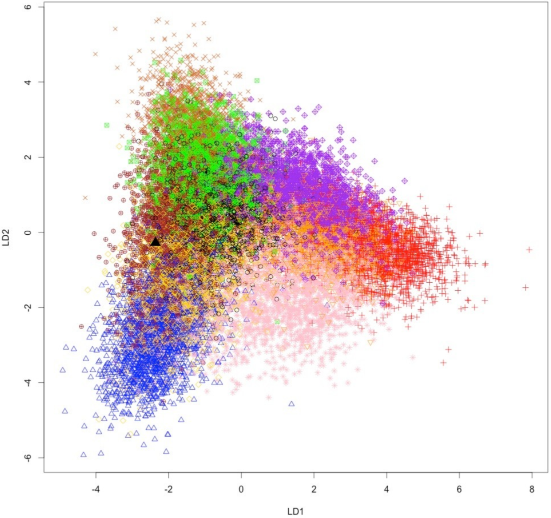
Evolution scenarios based on microsatellites

Posterior predictive error rates



favourable: 0.183 error – unfavourable: 0.435 error

Comparing 10 scenarios of Asian beetle invasion ◂ beetle moves



Back to Asian Ladybirds [message in a beetle]

Comparing 10 scenarios of Asian beetle invasion ◂ beetle moves

classification method	**prior error**[†] **rate** (in %)
raw LDA	38.94
"naïve" Bayes (with G margins)	54.02
k-nn (MAD normalised sum stat)	58.47
RF without LDA components	38.84
RF with LDA components	35.32

[†] estimated on pseudo-samples of 10^4 items drawn from the prior

Back to Asian Ladybirds [message in a beetle]

Comparing 10 scenarios of Asian beetle invasion ◂ beetle moves

Random forest allocation frequencies

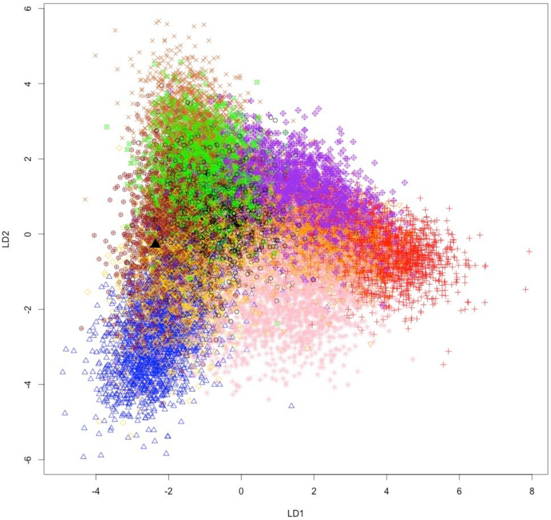
```
1     2   3     4     5     6     7     8    9     10
0.168 0.1 0.008 0.066 0.296 0.016 0.092 0.04 0.014 0.2
```

Posterior predictive error based on 20,000 prior simulations and keeping 500 neighbours (or 100 neighbours and 10 pseudo-datasets per parameter)

```
0.3682
```

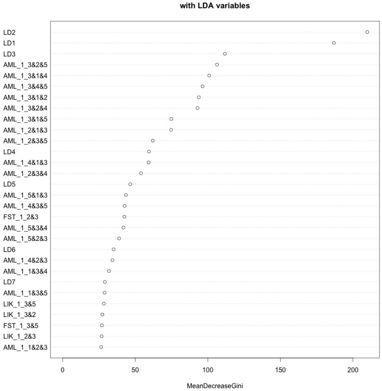
Back to Asian Ladybirds [message in a beetle]

Comparing 10 scenarios of Asian beetle invasion



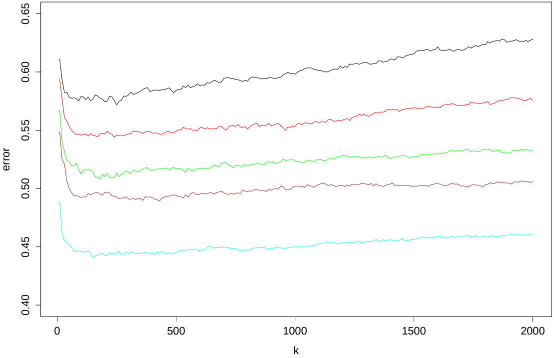
Back to Asian Ladybirds [message in a beetle]

Comparing 10 scenarios of Asian beetle invasion



Back to Asian Ladybirds [message in a beetle]

Comparing 10 scenarios of Asian beetle invasion



posterior predictive error 0.368

conclusion on random forests

- unlimited aggregation of arbitrary summary statistics
- recovery of discriminant statistics when available
- automated implementation with reduced calibration
- self-evaluation by posterior predictive error
- soon to appear in DIYABC

ABC estimation via random forests

1. simulation-based methods in Econometrics
2. Genetics of ABC
3. Approximate Bayesian computation
4. ABC for model choice
5. ABC model choice via random forests
6. **ABC estimation via random forests**
7. [some] asymptotics of ABC

Two basic issues with ABC

ABC compares numerous simulated dataset to the observed one

Two major difficulties:

- to decrease approximation error (or tolerance ϵ) and hence ensure reliability of ABC, total number of simulations very large;
- calibration of ABC (tolerance, distance, summary statistics, post-processing, &tc) critical and hard to automatise

classification of summaries by random forests

Given a large collection of summary statistics, rather than selecting a subset and excluding the others, estimate each parameter of interest by a machine learning tool like random forests

- RF can handle thousands of predictors
- ignore useless components
- fast estimation method with good local properties
- automatised method with few calibration steps
- substitute to Fearnhead and Prangle (2012) preliminary estimation of $\hat{\theta}(y^{\text{obs}})$
- includes a natural (classification) distance measure that avoids choice of both distance and tolerance

[Marin et al., 2016]

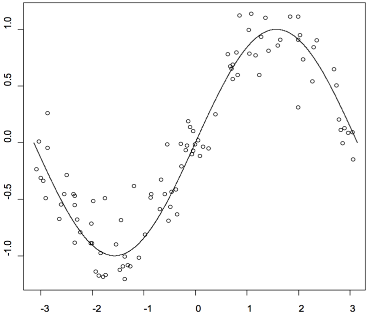
random forests as non-parametric regression

CART means Classification and Regression Trees

For regression purposes, i.e., to predict y as $f(x)$, similar binary trees in random forests

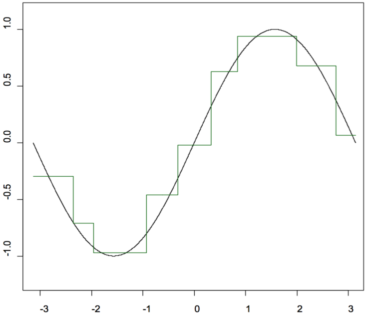
1. at each tree node, split data into two daughter nodes
2. split variable and bound chosen to minimise heterogeneity criterion
3. stop splitting when enough homogeneity in current branch
4. predicted values at terminal nodes (or leaves) are average response variable y for all observations in final leaf

Illustration



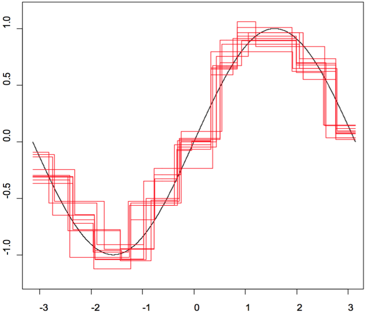
conditional expectation $f(x)$ and well-specified dataset

Illustration



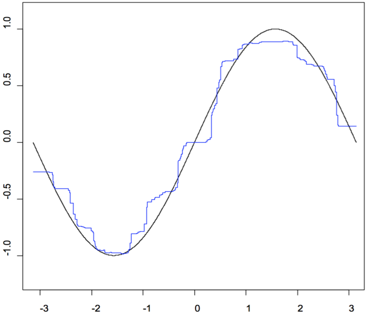
single regression tree

Illustration



ten regression trees obtained by bagging (Bootstrap AGGregatING)

Illustration



average of 100 regression trees

bagging reduces learning variance

When growing forest with many trees,

- grow each tree on an independent bootstrap sample
- at each node, select m variables at random out of all M possible variables
- Find the best dichotomous split on the selected m variables
- predictor function estimated by averaging trees

bagging reduces learning variance

When growing forest with many trees,

- grow each tree on an independent bootstrap sample
- at each node, select m variables at random out of all M possible variables
- Find the best dichotomous split on the selected m variables
- predictor function estimated by averaging trees

Improve on CART with respect to accuracy and stability

prediction error

A given simulation $(y^{\mathsf{sim}}, x^{\mathsf{sim}})$ in the training table is not used in about $^1/_3$ of the trees ("out-of-bag" case)

Average predictions $\hat{F}^{\mathsf{oob}}(x^{\mathsf{sim}})$ of these trees to give out-of-bag predictor of y^{sim}

Related methods

- adjusted local linear: Beaumont et al. (2002) Approximate Bayesian computation in population genetics, *Genetics*
- ridge regression: Blum et al. (2013) A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation, *Statistical Science*
- linear discriminant analysis: Estoup et al. (2012) Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics, *Molecular Ecology Resources*
- adjusted neural networks: Blum and François (2010) Non-linear regression models for Approximate Bayesian Computation, *Statistics and Computing*

ABC parameter estimation (ODOF)

One dimension = one forest (ODOF) methodology

ABC parameter estimation (ODOF)

One dimension = one forest (ODOF) methodology

parametric statistical model:

$$\{f(y;\theta)\colon y \in \mathcal{Y}, \theta \in \Theta\}, \quad \mathcal{Y} \subseteq \mathbb{R}^n, \quad \Theta \subseteq \mathbb{R}^p$$

with intractable density $f(\cdot;\theta)$

plus prior distribution $\pi(\theta)$

Inference on quantity of interest

$$\psi(\theta) \in \mathbb{R}$$

(posterior means, variances, quantiles or covariances)

common reference table

Given $\eta\colon \mathcal{Y} \to \mathbb{R}^k$ a collection of summary statistics

- produce reference table (RT) used as learning dataset for multiple random forests
- meaning, for $1 \leq t \leq N$
    1. simulate $\theta^{(t)} \sim \pi(\theta)$
    2. simulate $\tilde{y}_t = (\tilde{y}_{1,t}, \dots, \tilde{y}_{n,t}) \sim f(y; \theta^{(t)})$
    3. compute $\eta(\tilde{y}_t) = \{\eta_1(\tilde{y}_t), \dots, \eta_k(\tilde{y}_t)\}$

ABC posterior expectations

Recall that $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$

ABC posterior expectations

For each θ_j, construct a separate RF regression with predictors variables equal to summary statistics $\eta(y) = \{\eta_1(y), \ldots, \eta_k(y)\}$

If $L_b(\eta(y^*))$ denotes leaf index of b-th tree associated with $\eta(y^*)$ —leaf reached through path of binary choices in tree—, with $|L_b|$ response variables

$$\widehat{\mathbb{E}(\theta_j \mid \eta(y^*))} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{|L_b(\eta(y^*))|} \sum_{t:\eta(y_t)\in L_b(\eta(y^*))} \theta_j^{(t)}$$

is our ABC estimate

ABC posterior quantile estimate

Random forests also available for quantile regression

[Meinshausen, 2006, JMLR]

Since

$$\hat{\mathbb{E}}(\theta_j \mid \eta(y^*)) = \sum_{t=1}^{N} w_t(\eta(y^*))\theta_j^{(t)}$$

with

$$w_t(\eta(y^*)) = \frac{1}{B}\sum_{b=1}^{B} \frac{\mathbb{I}_{L_b(\eta(y^*))}(\eta(y_t))}{|L_b(\eta(y^*))|}$$

natural estimate of the cdf of θ_j is

$$\hat{F}(u \mid \eta(y^*)) = \sum_{t=1}^{N} w_t(\eta(y^*))\mathbb{I}_{\{\theta_j^{(t)} \leq u\}}.$$

ABC posterior quantile estimate

Since

$$\hat{\mathbb{E}}(\theta_j \mid \eta(y^*)) = \sum_{t=1}^{N} w_t(\eta(y^*))\theta_j^{(t)}$$

with

$$w_t(\eta(y^*)) = \frac{1}{B}\sum_{b=1}^{B} \frac{\mathbb{I}_{L_b(\eta(y^*))}(\eta(y_t))}{|L_b(\eta(y^*))|}$$

natural estimate of the cdf of θ_j is

$$\hat{F}(u \mid \eta(y^*)) = \sum_{t=1}^{N} w_t(\eta(y^*))\mathbb{I}_{\{\theta_j^{(t)} \le u\}}.$$

ABC posterior quantiles + credible intervals given by $\hat{F}^{-1}$

ABC variances

Even though approximation of $\mathrm{Var}(\theta_j \mid \eta(y^*))$ available based on $\hat{F}$, choice of alternative and slightly more involved version

In a given tree b in a random forest, existence of out-of-baf entries, i.e., not sampled in associated bootstrap subsample

Use of out-of-bag simulations to produce estimate of $\mathbb{E}\{\theta_j \mid \eta(y_t)\}$, $\tilde{\theta}_j{}^{(t)}$,

ABC variances

Even though approximation of $\mathrm{Var}(\theta_j \mid \eta(y^*))$ available based on $\hat{F}$, choice of alternative and slightly more involved version

In a given tree b in a random forest, existence of out-of-baf entries, i.e., not sampled in associated bootstrap subsample

Use of out-of-bag simulations to produce estimate of $\mathbb{E}\{\theta_j \mid \eta(y_t)\}$, $\tilde{\theta}_j^{(t)}$,

Apply weights $\omega_t(\eta(y^*))$ to out-of-bag residuals:

$$\widehat{\mathrm{Var}}(\theta_j \mid \eta(y^*)) = \sum_{t=1}^{N} \omega_t(\eta(y^*)) \left\{ (\theta_j^{(t)} - \tilde{\theta}_j^{(t)} \right\}^2$$

ABC covariances

For estimating $\mathrm{Cov}(\theta_j, \theta_\ell \mid \eta(y^*))$, construction of a specific random forest

product of out-of-bag errors for θ_j and θ_ℓ

$$\left\{\theta_j^{(t)} - \tilde{\theta}_j^{(t)}\right\}\left\{\theta_\ell^{(t)} - \tilde{\theta}_\ell^{(t)}\right\}$$

with again predictors variables the summary statistics $\eta(y) = \{\eta_1(y), \ldots, \eta_k(y)\}$

Gaussian toy example

Take

$$\begin{aligned}(y_1, \dots, y_n) \mid \theta_1, \theta_2 &\sim_{\text{iid}} \mathcal{N}(\theta_1, \theta_2), \quad n = 10\\ \theta_1 \mid \theta_2 &\sim \mathcal{N}(0, \theta_2)\\ \theta_2 &\sim I\mathcal{G}(4, 3)\\ \theta_1 \mid y &\sim \mathcal{T}\left(n+8, (n\bar{y})/(n+1), (s^2+6)/((n+1)(n+8))\right)\\ \theta_2 \mid y &\sim I\mathcal{G}\left\{n/2+4, s^2/2+3\right\}\end{aligned}$$

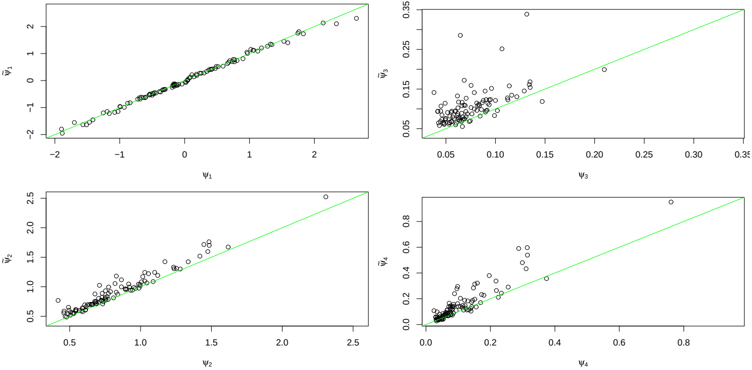
Closed-form theoretical values like
$\psi_1(y) = \mathbb{E}(\theta_1 \mid y)$, $\psi_2(y) = \mathbb{E}(\theta_2 \mid y)$, $\psi_3(y) = \text{Var}(\theta_1 \mid y)$ and
$\psi_4(y) = \text{Var}(\theta_2 \mid y)$

Gaussian toy example

Reference table of $N = 10,000$ Gaussian replicates
Independent Gaussian test set of size $N_{\mathrm{pred}} = 100$

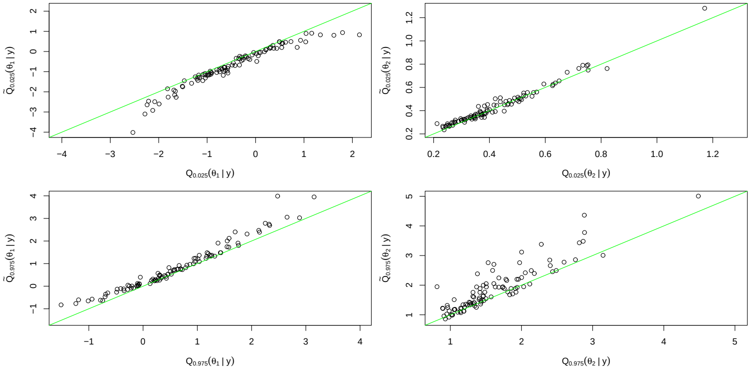
$k = 53$ **summary statistics**: the sample mean, the sample variance and the sample median absolute deviation, and 50 independent pure-noise variables (uniform [0,1])

Gaussian toy example



Scatterplot of the theoretical values with their corresponding estimates

Gaussian toy example



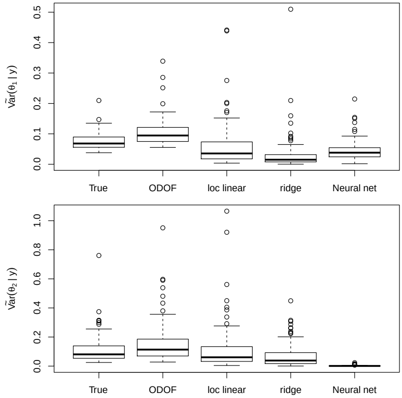
Scatterplot of the theoretical values of 2.5% and 97.5% posterior quantiles for θ_1 and θ_2 with their corresponding estimates

Gaussian toy example

	ODOF	adj local linear	adj ridge	adj neural net
$\psi_1(y) = \mathbb{E}(\theta_1 \mid y)$	**0.21**	0.42	0.38	0.42
$\psi_2(y) = \mathbb{E}(\theta_2 \mid y)$	**0.11**	0.20	0.26	0.22
$\psi_3(y) = \mathsf{Var}(\theta_1 \mid y)$	**0.47**	0.66	0.75	0.48
$\psi_4(y) = \mathsf{Var}(\theta_2 \mid y)$	**0.46**	0.85	0.73	0.98
$Q_{0.025}(\theta_1 \| y)$	0.69	0.55	0.78	**0.53**
$Q_{0.025}(\theta_2 \| y)$	**0.06**	0.45	0.68	1.02
$Q_{0.975}(\theta_1 \| y)$	**0.48**	0.55	0.79	0.50
$Q_{0.975}(\theta_2 \| y)$	**0.18**	0.23	0.23	0.38

Comparison of normalized mean absolute errors

Gaussian toy example



Boxplot comparison of $\mathrm{Var}(\theta_1 \mid y)$, $\mathrm{Var}(\theta_2 \mid y)$ with the true values, ODOF and usual ABC methods

Comments

ABC RF methods mostly insensitive both to strong correlations between the summary statistics and to the presence of noisy variables.

implies less number of simulations and no calibration

Next steps: adaptive schemes, deep learning, inclusion in DIYABC

[some] asymptotics of ABC

1. simulation-based methods in Econometrics
2. Genetics of ABC
3. Approximate Bayesian computation
4. ABC for model choice
5. ABC model choice via random forests
6. ABC estimation via random forests
7. [some] asymptotics of ABC

consistency of ABC posteriors

Asymptotic study of the ABC-posterior $\mathbf{z} = \mathbf{z}^{(n)}$

- ABC posterior consistency and convergence rate (in n)
- Asymptotic shape of $\pi_\epsilon(\cdot|\mathbf{y}^{(n)})$
- Asymptotic behaviour of $\hat{\theta}_\epsilon = \mathbb{E}^{\text{ABC}}[\boldsymbol{\theta}|\mathbf{y}^{(n)}]$

[Frazier et al., 2016]

consistency of ABC posteriors

- Concentration around true value and Bayesian consistency less stringent conditions on the convergence speed of tolerance ϵ_n to zero, when compared with asymptotic normality of ABC posterior
- asymptotic normality of ABC posterior mean does not require asymptotic normality of ABC posterior

ABC posterior consistency

For a sample $\mathbf{y} = \mathbf{y}^{(n)}$ and a tolerance $\epsilon = \epsilon_n$, when $n \to +\infty$, assuming a parametric model $\boldsymbol{\theta} \in \mathbb{R}^k$, k fixed

- Concentration of summary $\eta(\mathbf{z})$: there exists $b(\boldsymbol{\theta})$ such that

$$\eta(\mathbf{z}) - b(\boldsymbol{\theta}) = o_{\mathbb{P}_{\boldsymbol{\theta}}}(1)$$

- Consistency:

$$\Pi_{\epsilon_n}\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta | \mathbf{y}\right) = 1 + o_p(1)$$

- Convergence rate: there exists $\delta_n = o(1)$ such that

$$\Pi_{\epsilon_n}\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_n | \mathbf{y}\right) = 1 + o_p(1)$$

Related results

existing studies on the large sample properties of ABC, in which the asymptotic properties of point estimators derived from ABC have been the primary focus

[Creel et al., 2015; Jasra, 2015; Li & Fearnhead, 2015]

Convergence when $\epsilon_n \gtrsim \sigma_n$

Under assumptions

(A1) $\exists \sigma_n \to +\infty$

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\sigma_n^{-1}\|\eta(\mathbf{z}) - b(\boldsymbol{\theta})\| > u\right) \leq c(\boldsymbol{\theta})h(u), \quad \lim_{u\to+\infty} h(u) = 0$$

(A2)

$$\Pi(\|b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0)\| \leq u) \asymp u^D, \quad u \approx 0$$

posterior consistency and posterior concentration rate λ_T that depends on the deviation control of $d_2\{\eta(z), b(\theta)\}$
posterior concentration rate for $b(\theta)$ bounded from below by $O(\epsilon_T)$

Convergence when $\epsilon_n \gtrsim \sigma_n$

Under assumptions

(A1) $\exists \sigma_n \to +\infty$

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\sigma_n^{-1}\|\eta(\mathbf{z}) - b(\boldsymbol{\theta})\| > u\right) \leq c(\boldsymbol{\theta})h(u), \quad \lim_{u \to +\infty} h(u) = 0$$

(A2)

$$\Pi(\|b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0)\| \leq u) \asymp u^D, \quad u \approx 0$$

then

$$\Pi_{\epsilon_n}\left(\|b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0)\| \lesssim \epsilon_n + \sigma_n h^{-1}(\epsilon_n^D)|\mathbf{y}\right) = 1 + o_{p_0}(1)$$

If also $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq L\|b(\boldsymbol{\theta}) - c(\boldsymbol{\theta}_0)\|^\alpha$, locally and $\boldsymbol{\theta} \to b(\boldsymbol{\theta})$ 1-1

$$\Pi_{\epsilon_n}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \lesssim \underbrace{\epsilon_n^\alpha + \sigma_n^\alpha(h^{-1}(\epsilon_n^D))^\alpha}_{\delta_n}|\mathbf{y}) = 1 + o_{p_0}(1)$$

Comments

- if $\mathbb{P}_{\boldsymbol{\theta}}\left(\sigma_n^{-1}\|\eta(\mathbf{z}) - b(\boldsymbol{\theta})\| > u\right) \leq c(\boldsymbol{\theta})h(u)$, two cases
  1. Polynomial tail: $h(u) \lesssim u^{-\kappa}$, then $\delta_n = \epsilon_n + \sigma_n \epsilon_n^{-D/\kappa}$
  2. Exponential tail: $h(u) \lesssim e^{-cu}$, then $\delta_n = \epsilon_n + \sigma_n \log(1/\epsilon_n)$

Comments

- if $\mathbb{P}_{\boldsymbol{\theta}}\left(\sigma_n^{-1}\|\eta(\mathbf{z})-b(\boldsymbol{\theta})\|>u\right)\leq c(\boldsymbol{\theta})h(u)$, two cases
  1. Polynomial tail: $h(u)\lesssim u^{-\kappa}$, then $\delta_n=\epsilon_n+\sigma_n\epsilon_n^{-D/\kappa}$
  2. Exponential tail: $h(u)\lesssim e^{-cu}$, then $\delta_n=\epsilon_n+\sigma_n\log(1/\epsilon_n)$
- E.g., $\eta(\mathbf{y})=n^{-1}\sum_i g(y_i)$ with moments on g (case 1) or Laplace transform (case 2)

Comments

- $\Pi(\|b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0)\| \leq u) \asymp u^D$: If Π regular enough then $D = \mathrm{dim}(\theta)$
- no need to approximate the density $f(\eta(\mathbf{y})|\boldsymbol{\theta})$.
- Same results holds when $\epsilon_n = o(\sigma_n)$ if **(A2)** replaced with

$$\inf_{|x|\leq M} \mathbb{P}_{\boldsymbol{\theta}} \left(||\sigma_n^{-1}(\eta(\mathbf{z}) - b(\boldsymbol{\theta})) - x|| \leq u\right) \gtrsim u^D, \quad u \approx 0$$

proof

Simple enough proof: assume $\sigma_n \leq \delta\epsilon_n$ and

$$\|\eta(\mathbf{y}) - b(\boldsymbol{\theta}_0\|\| \lesssim \sigma_n, \quad \|\eta(\mathbf{y}) - \eta(\mathbf{z})\| \leq \epsilon_n$$

Hence

$$\|b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0)\| > \delta_n \quad \Rightarrow \quad \|\eta(\mathbf{z}) - b(\boldsymbol{\theta})\| > \delta_n - \epsilon_n - \sigma_n := t_n$$

proof

Simple enough proof: assume $\sigma_n \leq \delta\epsilon_n$ and

$$\|\eta(\mathbf{y}) - b(\boldsymbol{\theta}_0\|| \lesssim \sigma_n, \quad \|\eta(\mathbf{y}) - \eta(\mathbf{z})\| \leq \epsilon_n$$

Hence

$$\|b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0)\| > \delta_n \quad \Rightarrow \quad \|\eta(\mathbf{z}) - b(\boldsymbol{\theta})\| > \delta_n - \epsilon_n - \sigma_n := t_n$$

Also, if $\|b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0)\| \leq \epsilon_n/3$

$$\|\eta(\mathbf{y}) - \eta(\mathbf{z})\| \leq \|\eta(\mathbf{z}) - b(\boldsymbol{\theta})\| + \underbrace{\sigma_n}_{\leq \epsilon_n/3} + \epsilon_n/3$$

and

$$\Pi_{\epsilon_n}\left(\|b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0)\| > \delta_n|\mathbf{y}\right) \leq \frac{\int_{\|b(\boldsymbol{\theta})-b(\boldsymbol{\theta}_0)\|>\delta_n} \mathbb{P}_\theta\left(\|\eta(\mathbf{z}) - b(\boldsymbol{\theta})\| > t_n\right) d\Pi(\theta)}{\int_{\|b(\boldsymbol{\theta})-b(\boldsymbol{\theta}_0)\|\leq\epsilon_n/3} \mathbb{P}_\theta\left(\|\eta(\mathbf{z}) - b(\theta)\| \leq \epsilon_n/3\right) d\Pi(\theta)}$$

$$\lesssim \epsilon_n^{-D} h(t_n \sigma_n^{-1}) \int_\Theta c(\theta) d\Pi(\theta)$$

Summary statistic and (in)consistency

Consider the moving average MA(2) model

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, \quad e_t \sim_{i.i.d.} \mathcal{N}(0,1)$$

and

$$-2 \leq \theta_1 \leq 2,\ \theta_1 + \theta_2 \geq -1, \theta_1 - \theta_2 \leq 1.$$

Summary statistic and (in)consistency

Consider the moving average MA(2) model

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, \quad e_t \sim_{i.i.d.} \mathcal{N}(0,1)$$

and

$$-2 \leq \theta_1 \leq 2,\ \theta_1 + \theta_2 \geq -1, \theta_1 - \theta_2 \leq 1.$$

summary statistics equal to sample autocovariances

$$\eta_j(\boldsymbol{y}) = T^{-1} \sum_{t=1+j}^{T} y_t y_{t-j} \qquad j = 0, 1$$

with

$$\eta_0(\boldsymbol{y}) \xrightarrow{P} \mathbb{E}[y_t^2] = 1 + (\theta_{01})^2 + (\theta_{02})^2 \quad \text{and} \quad \eta_1(\boldsymbol{y}) \xrightarrow{P} \mathbb{E}[y_t y_{t-1}] = \theta_{01}(1 + \theta_{02})$$

Summary statistic and (in)consistency

Consider the moving average MA(2) model

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, \quad e_t \sim_{i.i.d.} \mathcal{N}(0,1)$$

and

$$-2 \leq \theta_1 \leq 2,\ \theta_1 + \theta_2 \geq -1, \theta_1 - \theta_2 \leq 1.$$

summary statistics equal to sample autocovariances

$$\eta_j(\boldsymbol{y}) = T^{-1} \sum_{t=1+j}^{T} y_t y_{t-j} \qquad j = 0, 1$$

with

$$\eta_0(\boldsymbol{y}) \xrightarrow{P} \mathbb{E}[y_t^2] = 1 + (\theta_{01})^2 + (\theta_{02})^2 \quad \text{and} \quad \eta_1(\boldsymbol{y}) \xrightarrow{P} \mathbb{E}[y_t y_{t-1}] = \theta_{01}(1 + \theta_{02})$$

For ABC target $p_\varepsilon(\boldsymbol{\theta}|\boldsymbol{\eta}(\boldsymbol{y}))$ to be degenerate at $\boldsymbol{\theta}_0$

$$\mathbf{0} = \boldsymbol{b}(\boldsymbol{\theta}_0) - \boldsymbol{b}(\boldsymbol{\theta}) = \begin{pmatrix} 1 + (\theta_{01})^2 + (\theta_{02})^2 \\ \theta_{01}(1 + \theta_{02}) \end{pmatrix} - \begin{pmatrix} 1 + (\theta_1)^2 + (\theta_2)^2 \\ \theta_1(1 + \theta_2) \end{pmatrix}$$

must have unique solution $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

Summary statistic and (in)consistency

Consider the moving average MA(2) model

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, \quad e_t \sim_{i.i.d.} \mathcal{N}(0,1)$$

and

$$-2 \leq \theta_1 \leq 2,\ \theta_1 + \theta_2 \geq -1, \theta_1 - \theta_2 \leq 1.$$

summary statistics equal to sample autocovariances

$$\eta_j(\boldsymbol{y}) = T^{-1} \sum_{t=1+j}^{T} y_t y_{t-j} \qquad j = 0, 1$$

with

$$\eta_0(\boldsymbol{y}) \overset{P}{\to} \mathbb{E}[y_t^2] = 1 + (\theta_{01})^2 + (\theta_{02})^2 \quad \text{and} \quad \eta_1(\boldsymbol{y}) \overset{P}{\to} \mathbb{E}[y_t y_{t-1}] = \theta_{01}(1 + \theta_{02})$$

Take $\theta_{01} = .6, \theta_{02} = .2$: equation has 2 solutions

$$\theta_1 = .6, \theta_2 = .2 \quad \text{and} \quad \theta_1 \approx .5453, \theta_2 \approx .3204$$

Asymptotic shape of posterior distribution

Three different regimes:

1. $\sigma_n = o(\epsilon_n) \longrightarrow$ Uniform limit
2. $\sigma_n \asymp \epsilon_n \longrightarrow$ perturbated Gaussian limit
3. $\sigma_n \gg \epsilon_n \longrightarrow$ Gaussian limit

Assumptions

- **(B1)** Concentration of summary η: $\Sigma_n(\boldsymbol{\theta}) \in \mathbb{R}^{k_1 \times k_1}$ is $o(1)$

  $$\Sigma_n(\boldsymbol{\theta})^{-1}\{\eta(\mathbf{z}) - b(\boldsymbol{\theta})\} \Rightarrow \mathcal{N}_{k_1}(0, Id), \quad (\Sigma_n(\theta)\Sigma_n(\theta_0)^{-1})_n = \mathcal{C}^o$$

- **(B2)** $b(\boldsymbol{\theta})$ is $\mathcal{C}^1$ and

  $$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \lesssim \|b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0)\|$$

- **(B3)** Dominated convergence and

  $$\lim_n \frac{\mathbb{P}_\theta(\Sigma_n(\theta)^{-1}\{\eta(\mathbf{z}) - b(\boldsymbol{\theta})\} \in u + B(0, \mathbf{u_n}))}{\prod_j u_n(j)} \rightarrow \varphi(u)$$

main result

Set $\Sigma_n(\boldsymbol{\theta}) = \sigma_n D(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \approx \boldsymbol{\theta}_0$ and $Z^o = \Sigma_n(\boldsymbol{\theta}_0)^{-1}(\eta(\mathbf{y}) - \boldsymbol{\theta}_0)$, then under **(B1)** and **(B2)**

- when $\epsilon_n \sigma_n^{-1} \to +\infty$

$$\Pi_{\epsilon_n}[\epsilon_n^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in A|\mathbf{y}] \Rightarrow \mathcal{U}_{B_0}(A), \quad B_0 = \{x \in \mathbb{R}^k ; \|b'(\boldsymbol{\theta}_0)^T x\| \leq 1\}$$

- when $\epsilon_n \sigma_n^{-1} \to c$

$$\Pi_{\epsilon_n}[\Sigma_n(\boldsymbol{\theta}_0)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - Z^o \in A|\mathbf{y}] \Rightarrow Q_c(A), \quad Q_c \neq \mathcal{N}$$

- when $\epsilon_n \sigma_n^{-1} \to 0$ and **(B3)** holds, set

$$V_n = [b'(\boldsymbol{\theta}_0)]^T \Sigma_n(\boldsymbol{\theta}_0) b'(\boldsymbol{\theta}_0)$$

then

$$\Pi_{\epsilon_n}[V_n^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \tilde{Z}^o \in A|\mathbf{y}] \Rightarrow \Phi(A),$$

intuition

Set $x(\theta) = \sigma_n^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - Z^o \ (k = 1)$

$$\pi_n := \Pi_{\epsilon_n}[\epsilon_n^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in A|\mathbf{y}]$$

$$= \frac{\int_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0| \le u_n} \mathbb{I}_{x(\boldsymbol{\theta}) \in A} \mathbb{P}_{\boldsymbol{\theta}} \left(\|\sigma_n^{-1}(\eta(\mathbf{z}) - b(\boldsymbol{\theta})) + x(\boldsymbol{\theta})\| \le \sigma_n^{-1}\epsilon_n \right) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0| \le u_n} \mathbb{P}_{\boldsymbol{\theta}} \left(\|\sigma_n^{-1}(\eta(\mathbf{z}) - b(\boldsymbol{\theta})) + x(\boldsymbol{\theta})\| \le \sigma_n^{-1}\epsilon_n \right) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} + o_p(1)$$

- If $\epsilon_n/\sigma_n \gg 1$:

  $$\mathbb{P}_{\boldsymbol{\theta}} \left(|\sigma_n^{-1}(\eta(\mathbf{z}) - b(\boldsymbol{\theta})) + x(\boldsymbol{\theta})| \le \sigma_n^{-1}\epsilon_n \right) = 1 + o(1), \text{ iff } \quad |x| \le \sigma_n^{-1}\epsilon$$

- If $\epsilon_n/\sigma_n = o(1)$

  $$\mathbb{P}_{\boldsymbol{\theta}} \left(|\sigma_n^{-1}(\eta(\mathbf{z}) - b(\boldsymbol{\theta})) + x| \le \sigma_n^{-1}\epsilon_n \right) = \phi(x)\sigma_n(1 + o(1))$$

more comments

- Surprising : $\mathcal{U}(-\epsilon_n, \epsilon_n)$ limit when $\epsilon_n \gg \sigma_n$

more comments

- Surprising : $\mathcal{U}(-\epsilon_n, \epsilon_n)$ limit when $\epsilon_n \gg \sigma_n$ but not so much since $\epsilon_n = o(1)$ means concentration around $\boldsymbol{\theta}_0$ and $\sigma_n = o(\epsilon_n)$ implies that $b(\boldsymbol{\theta}) - b(\boldsymbol{\theta}_0) \approx \eta(\mathbf{z}) - \eta(\mathbf{y})$
- again, there is no true need to control approximation of $f(\eta(\mathbf{y})|\boldsymbol{\theta})$ by a Gaussian density: merely a control of the distribution
- we have

$$\tilde{Z}^o = Z^o \Big/ b'(\boldsymbol{\theta}_0) \text{ like the asym score}$$

- generalisation to the case where eigenvalues of Σ_n are $d_{n,1} \neq \cdots \neq d_{n,k}$
- behaviour of $\mathbb{E}_{ABC}(\boldsymbol{\theta}|\mathbf{y})$ as in Li & Fearnhead (2016)

even more comments

If (also) $p(\theta)$ is Hölder β

$$\mathbb{E}_{ABC}(\boldsymbol{\theta}|\mathbf{y}) - \boldsymbol{\theta}_0 = \underbrace{\sigma_n \frac{Z^o}{b(\boldsymbol{\theta}_0)'}}_{\text{score for } f(\eta(\mathbf{y})|\boldsymbol{\theta})} + \underbrace{\sum_{j=1}^{\lfloor \beta/2 \rfloor} \epsilon_n^{2j} H_j(\boldsymbol{\theta}_0, p, b)}_{\text{bias from threshold approx}} + o(\sigma_n) + O(\epsilon_n^{\beta+1})$$

- if $\epsilon_n^2 = o(\sigma_n)$: *Efficiency*

$$\mathbb{E}_{ABC}(\boldsymbol{\theta}|\mathbf{y}) - \boldsymbol{\theta}_0 = \sigma_n \frac{Z^o}{b(\boldsymbol{\theta}_0)'} + o(\sigma_n)$$

- the $H_j(\boldsymbol{\theta}_0, p, b)$'s are deterministic

we gain nothing by getting a first crude $\hat{\theta}(\mathbf{y}) = \mathbb{E}_{ABC}(\boldsymbol{\theta}|\mathbf{y})$ for some $\eta(\mathbf{y})$ and then rerun ABC with $\hat{\theta}(\mathbf{y})$

impact of the dimension of η

dimension of $\eta(.)$ does not impact above result, but impacts acceptance probability

- if $\epsilon_n = o(\sigma_n)$, $k_1 = \dim(\eta(\mathbf{y}))$, $k = \dim(\boldsymbol{\theta})$ & $k_1 \geq k$

$$\alpha_n := \Pr\left(\|\mathbf{y} - \mathbf{z}\| \leq \epsilon_n\right) \asymp \epsilon_n^{k_1} \sigma_n^{-k_1+k}$$

- if $\epsilon_n \gtrsim \sigma_n$

$$\alpha_n := \Pr\left(\|\mathbf{y} - \mathbf{z}\| \leq \epsilon_n\right) \asymp \epsilon_n^{k}$$

- If we choose α_n
  - $\alpha_n = o(\sigma_n^k)$ leads to $\epsilon_n = \sigma_n(\alpha_n \sigma_n^{-k})^{1/k_1} = o(\sigma_n)$
  - $\alpha_n \gtrsim \sigma_n$ leads to $\epsilon_n \asymp \alpha_n^{1/k}$.

conclusion on ABC consistency

- asymptotic description of ABC: different regimes depending on ϵ_n σ_n
- no point in choosing ϵ_n arbitrarily small: just $\epsilon_n = o(\sigma_n)$
- no gain in iterative ABC
- results under weak conditions by not studying $g(\eta(\mathbf{z})|\theta)$

conclusion on ABC consistency

- asymptotic description of ABC: different regimes depending on ϵ_n σ_n
- no point in choosing ϵ_n arbitrarily small: just $\epsilon_n = o(\sigma_n)$
- no gain in iterative ABC
- results under weak conditions by not studying $g(\eta(\mathbf{z})|\theta)$

the end