ABC Methods for Bayesian Model Choice

Christian P. Robert

Université Paris-Dauphine, IuF, & CREST
http://www.ceremade.dauphine.fr/~xian

Joint work(s) with Jean-Marie Cornuet, Aude Grelaud, Jean-Michel Marin, Natesh Pillai, & Judith Rousseau

Approximate Bayesian computation

Regular Bayesian computation issues

When faced with a non-standard posterior distribution

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

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

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

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

[Robert & Casella, 2004]

Untractable likelihoods

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

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

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

© MCMC cannot be implemented!

Untractable likelihoods

© MCMC cannot be implemented!

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, $\mathbf{z} = \mathbf{y}$.

[Rubin, 1984; Tavaré et al., 1997]

A as approximative

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 approximative

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

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

ABC algorithm

Algorithm 1 Likelihood-free rejection sampler

for $i = 1$ to N **do**
  repeat
    generate θ' 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 (maybe in-sufficient) statistic

Output

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

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

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

Output

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

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

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

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

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

[Not garanteed!]

ABC for model choice

Bayesian model choice

Principle

Several models

$$M_1, M_2, \ldots$$

are considered simultaneously for dataset $\mathbf{y}$ and model index $\mathcal{M}$ central to inference.
Use of a prior $\pi(\mathcal{M} = m)$, 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 $\mathcal{M}$,

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

a challenging computational target when models are complex.

Generic ABC for model choice

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

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

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

ABC estimates

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

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

Early 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 across models?

ABC estimates

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

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

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

[Cornuet et al., DIYABC, 2009]

Potts model

Potts model

Distribution with an energy function of the form

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

where $l \sim i$ denotes a neighbourhood structure

Potts model

Potts model

Distribution with an energy function of the form

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

where $l \sim i$ denotes a neighbourhood structure

In most realistic settings, summation

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

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

Neighbourhood relations

Setup

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

Computational target:

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

Model index

Computational target:

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

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

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

Sufficient statistics in Gibbs random fields

Sufficient statistics in Gibbs random fields

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

Sufficient statistics in Gibbs random fields

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

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

where

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

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

More about sufficiency

'Sufficient statistics for individual models are unlikely to be very informative for the model probability. This is already well known and understood by the ABC-user community.'

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

More about sufficiency

'Sufficient statistics for individual models are unlikely to be very informative for the model probability. This is already well known and understood by the ABC-user community.'

[Scott Sisson, Jan. 31, 2011, '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)

More about sufficiency

'Sufficient statistics for individual models are unlikely to be very informative for the model probability. This is already well known and understood by the ABC-user community.'

[Scott Sisson, Jan. 31, 2011, '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

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

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

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

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,

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

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

When ϵ goes to zero,

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

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

Limiting behaviour of B_{12} (under sufficiency)

If $\eta(\mathbf{y})$ sufficient statistic in 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 in 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 $\mathcal{G}(p)$
Sum

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

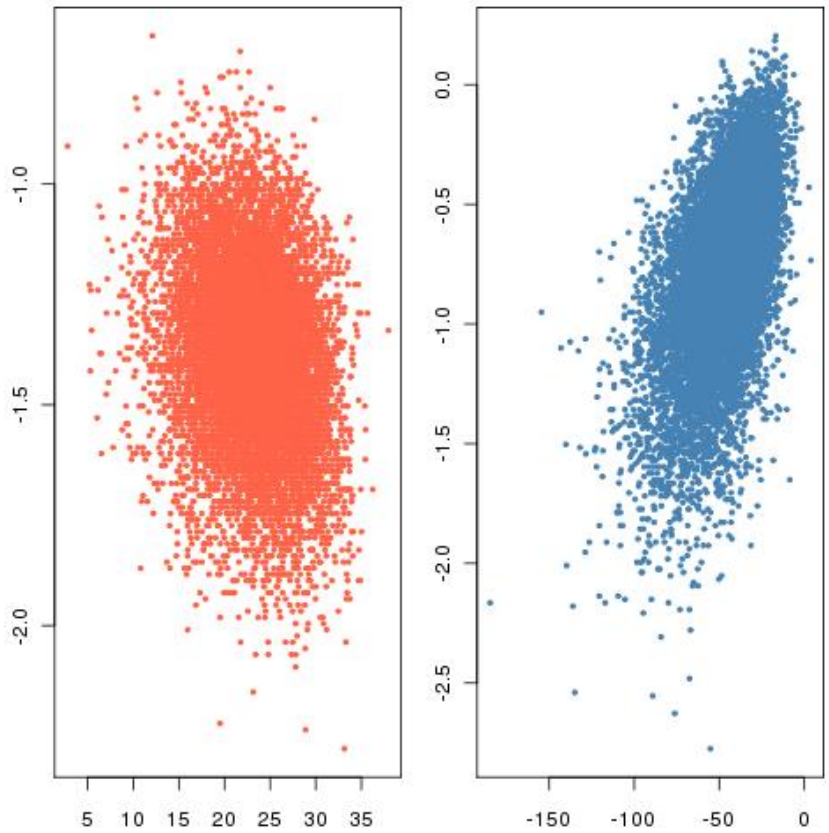
sufficient statistic for either model but not simultaneously

Discrepancy ratio

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

Poisson/geometric discrepancy

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



Formal recovery

Creating an encompassing exponential family

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

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

[Didelot, Everitt, Johansen & Lawson, 2011]

Formal recovery

Creating an encompassing exponential family

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

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

[Didelot, Everitt, Johansen & Lawson, 2011]

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

Formal recovery

Creating an encompassing exponential family

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

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

[Didelot, Everitt, Johansen & Lawson, 2011]

Only applies in genuine sufficiency settings...

© **Inability to evaluate 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, 'Og]

Meaning of the ABC-Bayes factor

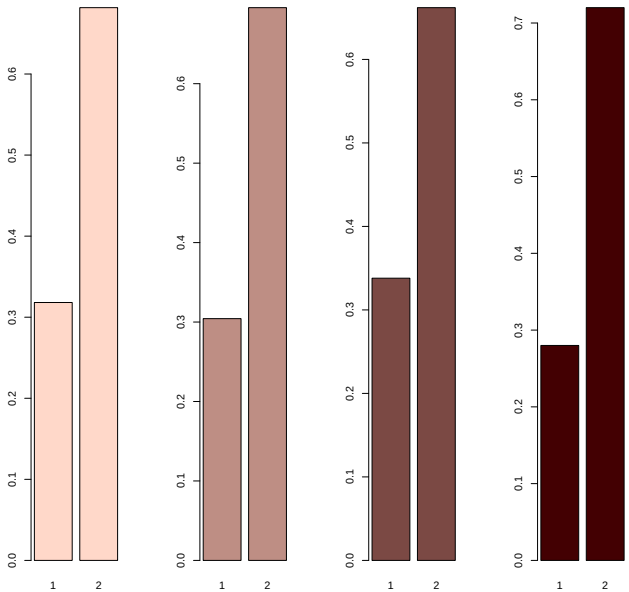
'This is also why focus on model discrimination typically (...) proceeds by (...) accepting that the Bayes Factor that one obtains is only derived from the summary statistics and may in no way correspond to that of the full model.'

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

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

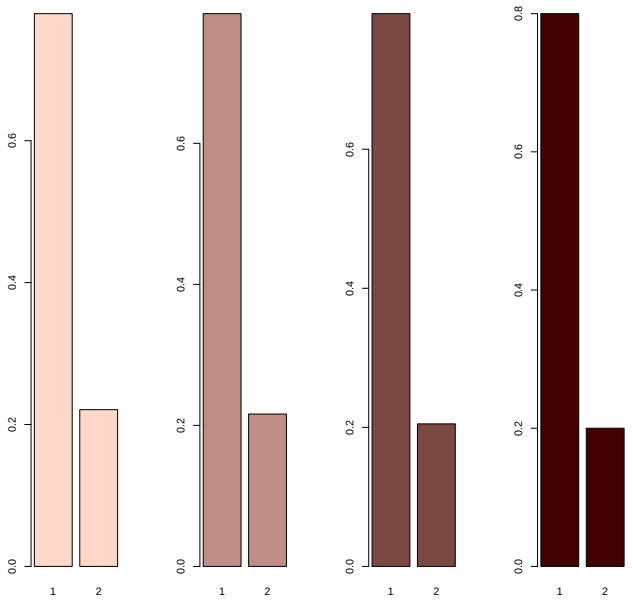
$$\lim_{n \to \infty} B_{12}^{\eta}(\mathbf{y}) = \frac{(\theta_0 + 1)^2}{\theta_0} e^{-\theta_0}$$

MA example



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

MA example



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

A population genetics evaluation

Population genetics example with

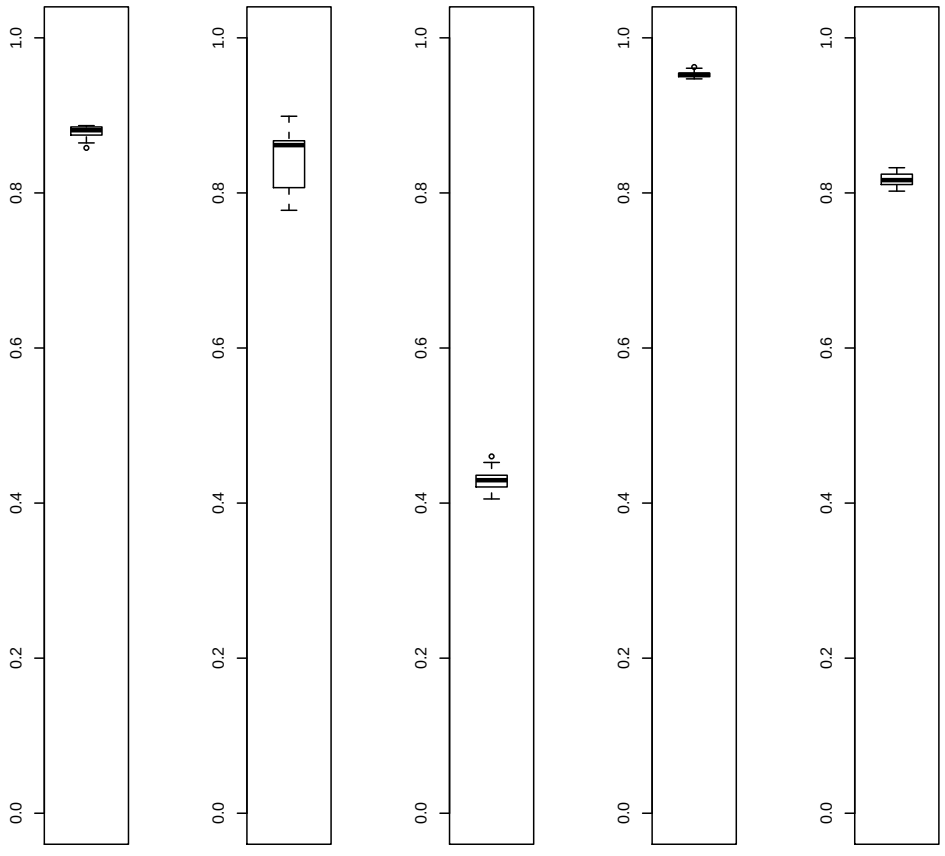
- 3 populations
- 2 scenari
- 15 individuals
- 5 loci
- single mutation parameter

A population genetics evaluation

Population genetics example with

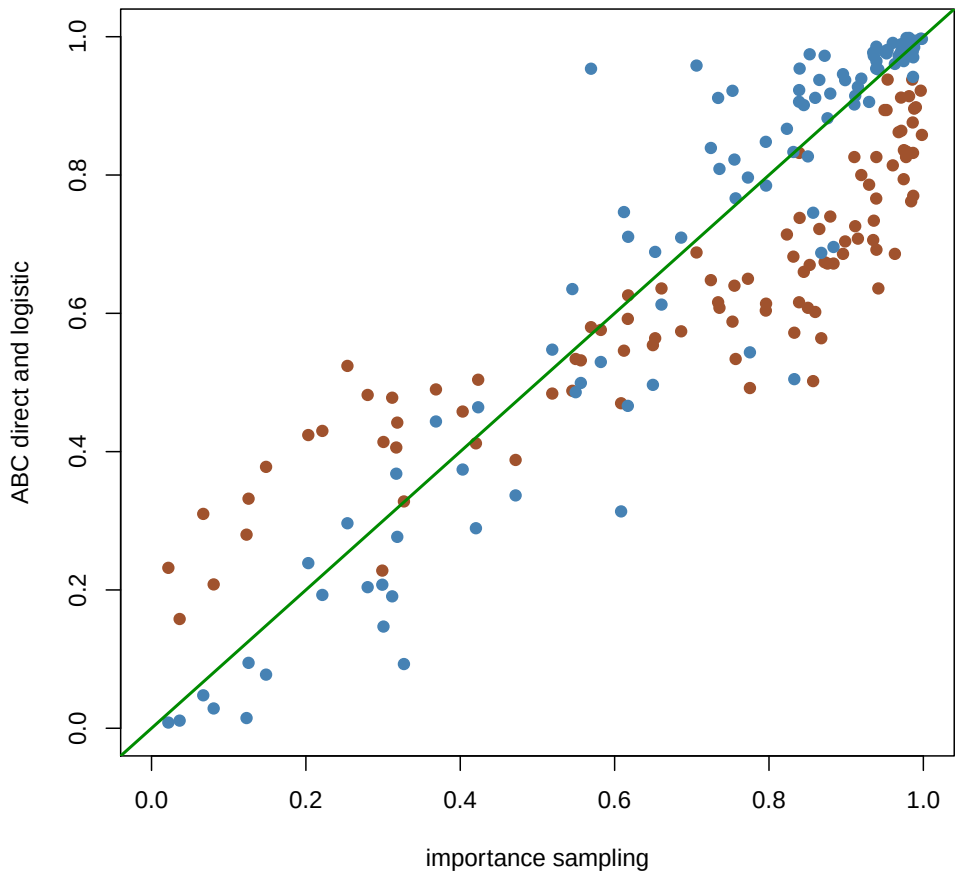
- 3 populations
- 2 scenari
- 15 individuals
- 5 loci
- single mutation parameter
- 24 summary statistics
- 2 million ABC proposal
- importance [tree] sampling alternative

Stability of importance sampling



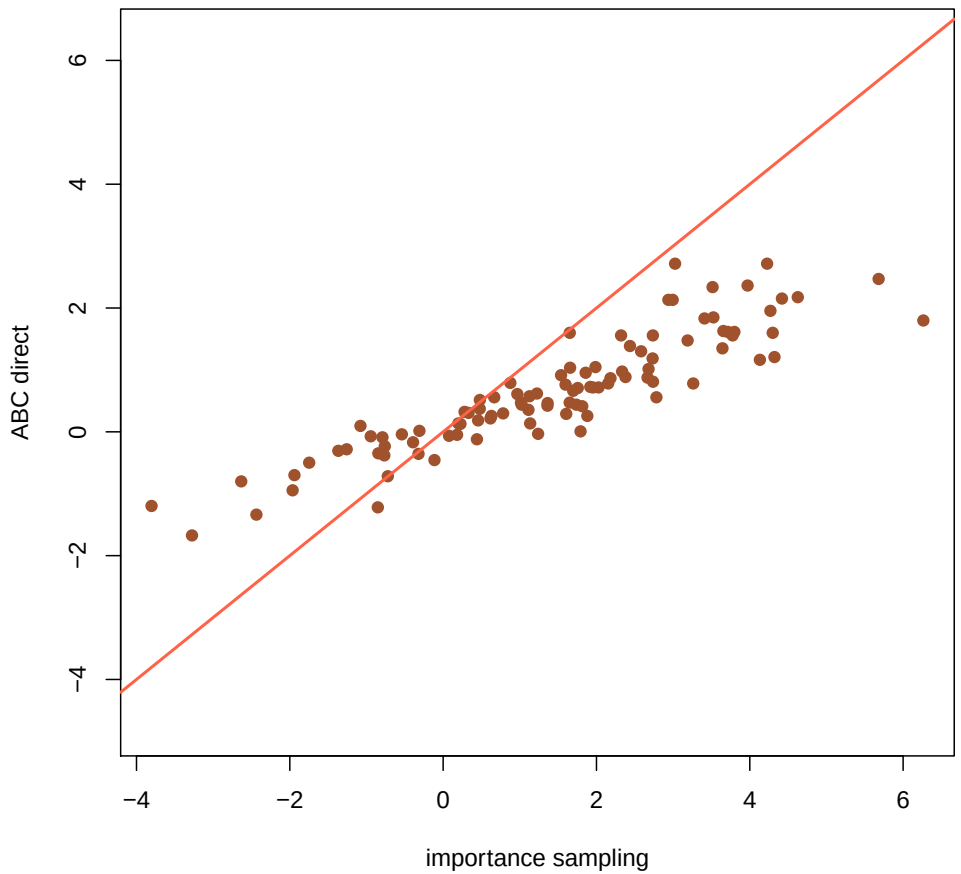
Comparison with ABC

Use of 24 summary statistics and DIY-ABC logistic correction



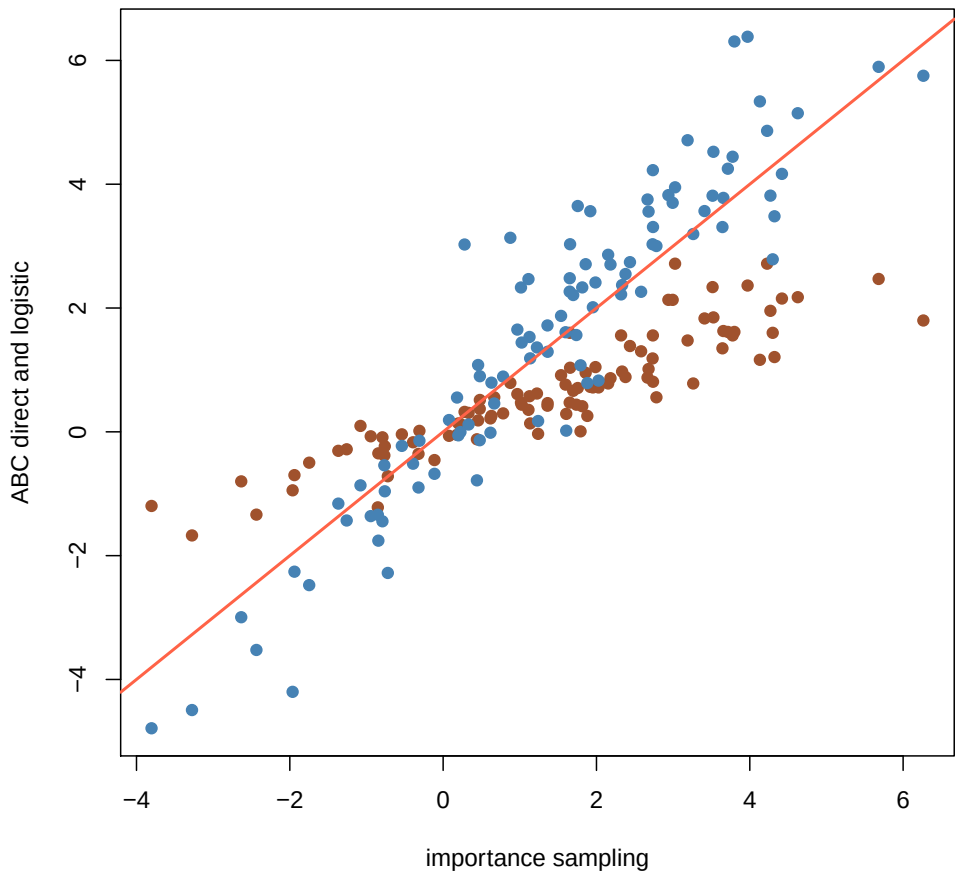
Comparison with ABC

Use of 15 summary statistics and DIY-ABC logistic correction



Comparison with ABC

Use of 15 summary statistics and DIY-ABC logistic correction



The only safe cases???

Besides specific models like Gibbs random fields,

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

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

The only safe cases???

Besides specific models like Gibbs random fields,

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

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

ABC model choice consistency

The starting point

Central question to the validation of ABC for model choice:

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

The starting point

Central question to the validation of ABC for model choice:

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

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

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks]:
[X, Cornuet, Marin, & Pillai, Aug. 2011]

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

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks]:
[X, Cornuet, Marin, & Pillai, Aug. 2011]

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

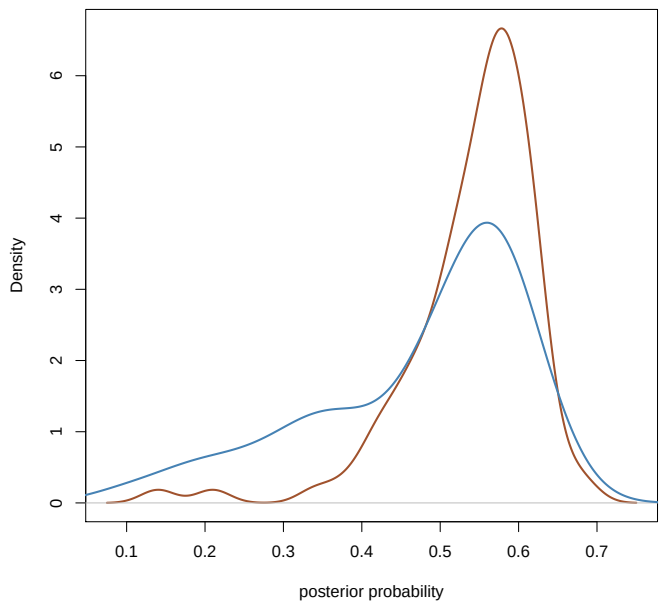
Four possible statistics

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

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks]:
[X, Cornuet, Marin, & Pillai, Aug. 2011]

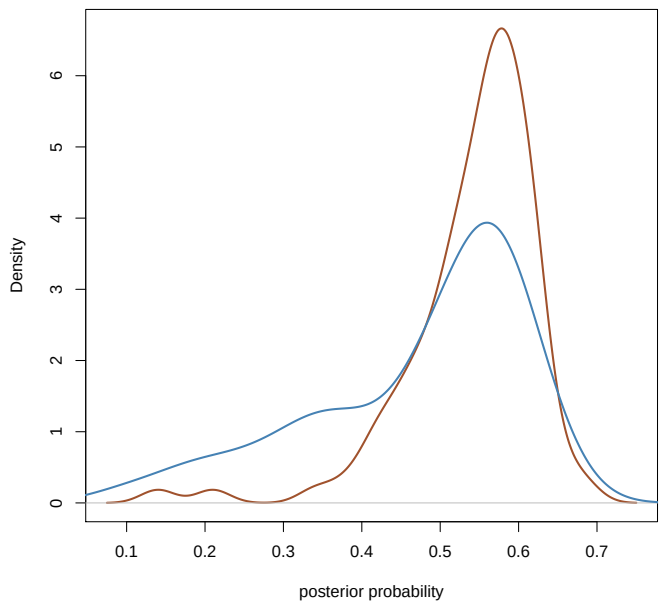
Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$:
$\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).

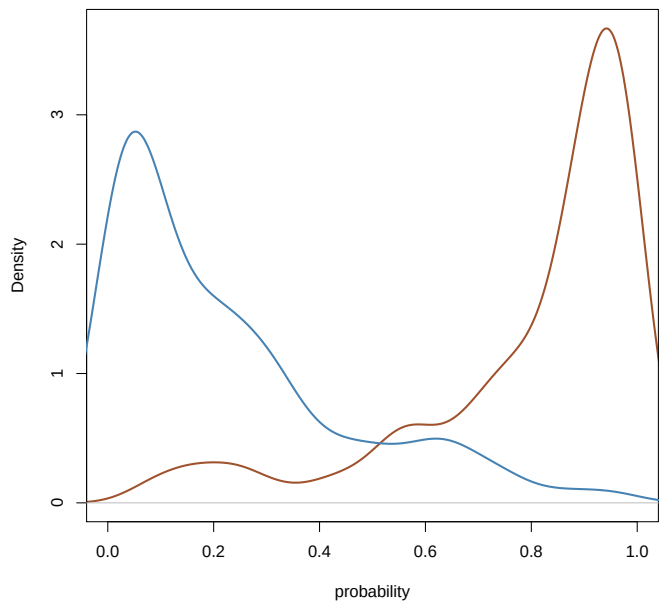


A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks]:
[X, Cornuet, Marin, & Pillai, Aug. 2011]

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





Framework

Starting from sample

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

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] There exist a sequence $\{v_n\}$ converging to $+\infty$,
an a.c. distribution Q with continuous bounded density $q(\cdot)$,
a symmetric, $d \times d$ positive definite matrix V_0
and a vector $\mu_0 \in \mathbb{R}^d$ such that

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

and for all $M > 0$

$$\sup_{v_n|t-\mu_0|<M} \left| |V_0|^{1/2} v_n^{-d} g_n(t) - q\big(v_n V_0^{-1/2}\{t-\mu_0\}\big) \right| = o(1)$$

Assumptions

A collection of asymptotic "standard" assumptions:

[A2] For $i = 1, 2$, there exist $d \times d$ symmetric positive definite matrices $V_i(\theta_i)$ and $\mu_i(\theta_i) \in \mathbb{R}^d$ such that

$$v_n V_i(\theta_i)^{-1/2}(\boldsymbol{T}^n - \mu_i(\theta_i)) \overset{n\to\infty}{\rightsquigarrow} Q, \qquad \text{under } G_{i,n}(\cdot|\theta_i) \ .$$

Assumptions

A collection of asymptotic "standard" assumptions:

[A3] For $i = 1, 2$, there exist sets $\mathcal{F}_{n,i} \subset \Theta_i$ and constants $\epsilon_i, \tau_i, \alpha_i > 0$ such that for all $\tau > 0$,

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

with

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

Assumptions

A collection of asymptotic "standard" assumptions:

[A4] For $u > 0$

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

if $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, there exist constants $d_i < \tau_i \wedge \alpha_i - 1$ such that

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

Assumptions

A collection of asymptotic "standard" assumptions:

[A5] If $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, there exists $U > 0$ such that for any $M > 0$,

$$\sup_{v_n|t-\mu_0|<M} \sup_{\theta_i \in S_{n,i}(U)} \Big| |V_i(\theta_i)|^{1/2} v_n^{-d} g_i(t|\theta_i) - q\big(v_n V_i(\theta_i)^{-1/2}(t - \mu(\theta_i))\big) \Big| = o(1)$$

and

$$\lim_{M\to\infty} \limsup_n \frac{\pi_i\Big(S_{n,i}(U) \cap \big\{\|V_i(\theta_i)^{-1}\| + \|V_i(\theta_i)\| > M\big\}\Big)}{\pi_i(S_{n,i}(U))} = 0 \,.$$

Assumptions

A collection of asymptotic "standard" assumptions:

[A1]–[A2] are standard central limit theorems (**[A1]** redundant when one model is "true")
[A3] controls the large deviations of the estimator $\boldsymbol{T}^n$ from the estimand $\mu(\theta)$
[A4] is the standard prior mass condition found in Bayesian asymptotics (d_i effective dimension of the parameter)
[A5] controls more tightly convergence esp. when μ_i is not one-to-one

Effective dimension

[A4] Understanding d_1, d_2 : defined only when $\mu_0 \in \{\mu_i(\theta_i), \theta_i \in \Theta_i\}$,

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

is the effective dimension of the model Θ_i around μ_0

Asymptotic marginals

Asymptotically, under **[A1]–[A5]**

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

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

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

Within-model consistency

Under same assumptions, if $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, the posterior distribution of $\mu_i(\theta_i)$ given $\boldsymbol{T}^n$ is consistent at rate $1/v_n$ provided $\alpha_i \wedge \tau_i > d_i$.

Within-model consistency

Under same assumptions, if $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, the posterior distribution of $\mu_i(\theta_i)$ given $\boldsymbol{T}^n$ is consistent at rate $1/v_n$ provided $\alpha_i \wedge \tau_i > d_i$.

Note: d_i can truly be seen as an effective dimension of the model under the posterior $\pi_i(.|\boldsymbol{T}^n)$, since if $\mu_0 \in \{\mu_i(\theta_i); \theta_i \in \Theta_i\}$,

$$m_i(\boldsymbol{T}^n) \sim v_n^{d-d_i}$$

Between-model consistency

Consequence of above is that asymptotic behaviour of the Bayes factor is driven by the asymptotic mean value 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 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 $d_1 - d_2$

Between-model consistency

Consequence of above is that asymptotic behaviour of the Bayes factor is driven by the asymptotic mean value 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

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

Consistency theorem

If

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

Bayes factor

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

irrespective of the true model. It is consistent iff P_n is within the model with the smallest dimension

Consistency theorem

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

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

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

Consequences on summary statistics

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

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

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

Toy example: Laplace versus Gauss [1]

If

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

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

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

Toy example: Laplace versus Gauss [1]

If

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

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

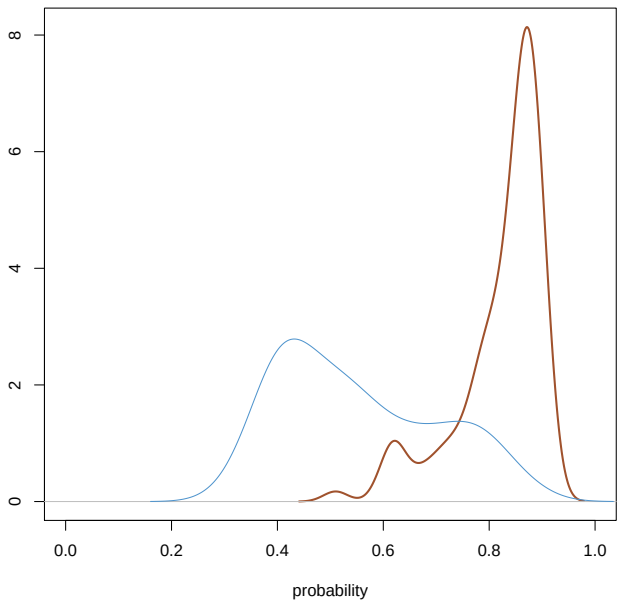
[here $d_1 = d_2 = d = 1$]

© a Bayes factor associated with such a statistic is inconsistent

Toy example: Laplace versus Gauss [1]

If

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

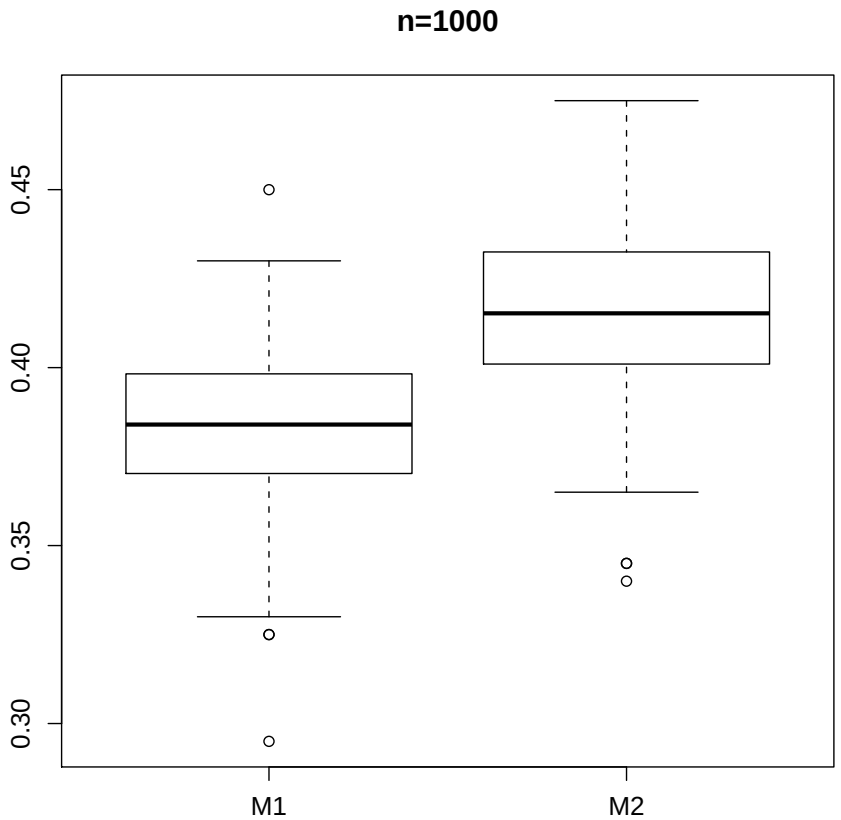


Fourth moment

Toy example: Laplace versus Gauss [1]

If

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



Toy example: Laplace versus Gauss [1]

If

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

Caption: *Comparison of the distributions of the posterior probabilities that the data is from a normal model (as opposed to a Laplace model) with unknown mean when the data is made of* $n = 1000$ *observations either from a normal (M1) or Laplace (M2) distribution with mean one and when the summary statistic in the ABC algorithm is restricted to the empirical fourth moment. The ABC algorithm uses proposals from the prior* $\mathcal{N}(0, 4)$ *and selects the tolerance as the 1% distance quantile.*

Toy example: Laplace versus Gauss [0]

When

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

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

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

thus

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

Toy example: Laplace versus Gauss [0]

When

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

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

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

thus

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

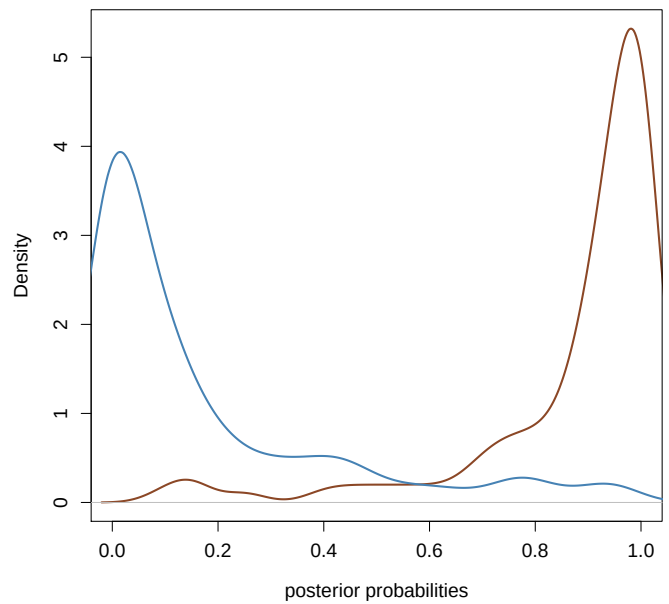
Under the Gaussian model $\mu_0 = 3$ $\mu_2(\theta_2) \geq 6 > 3$ $\forall \theta_2$

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

Toy example: Laplace versus Gauss [0]

When

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



Fourth AND sixth moments

Embedded models

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

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

Embedded models

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

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

If summary statistic only informative on a parameter that is the same under both models, i.e if $d_1 = d_2$, then
© the Bayes factor is not consistent

Embedded models

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

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

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

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

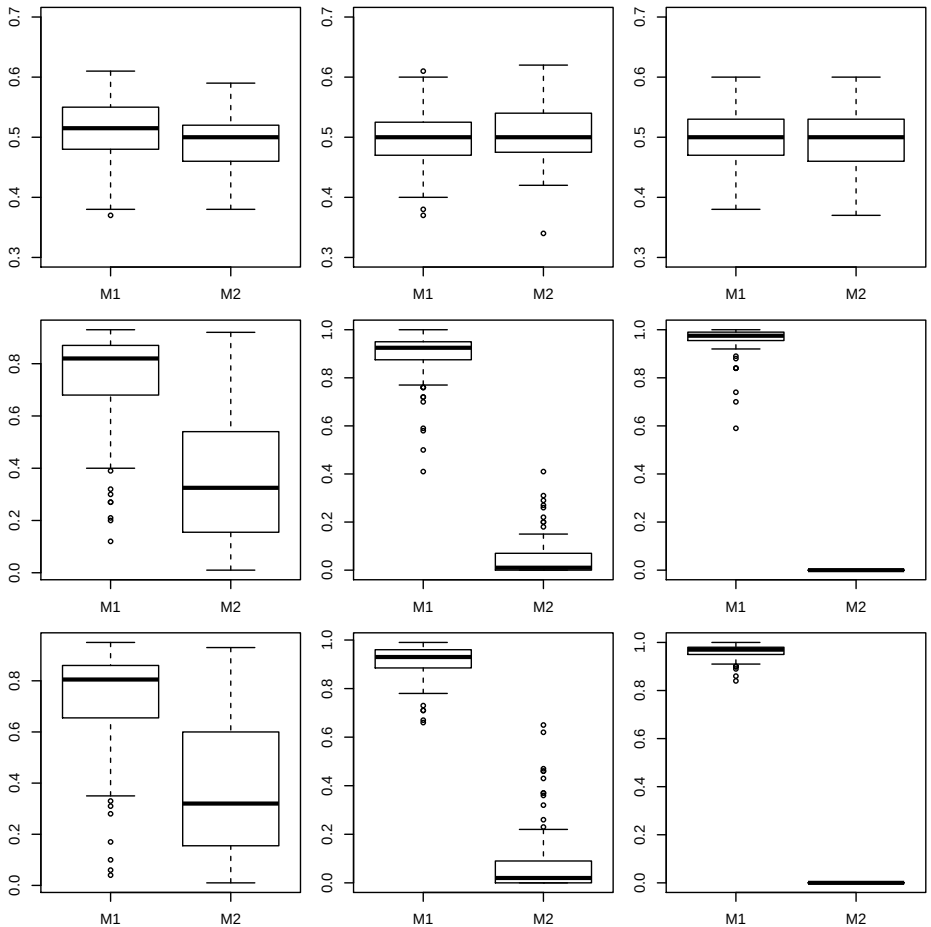
Another toy example: Quantile distribution

$$Q(p; A, B, g, k) = A + B \left[1 + \frac{1 - \exp\{-gz(p)\}}{1 + \exp\{-gz(p)\}}\right] \left[1 + z(p)^2\right]^k z(p)$$

A, B, g and k, location, scale, skewness and kurtosis parameters
Embedded models:

- $\mathfrak{M}_1$: $g = 0$ and $k \sim \mathcal{U}[-1/2, 5]$
- $\mathfrak{M}_2$: $g \sim \mathcal{U}[0, 4]$ and $k \sim \mathcal{U}[-1/2, 5]$.

Consistency [or not]



Consistency [or not]

Caption: *Comparison of the distributions of the posterior probabilities that the data is from model* $\mathfrak{M}_1$ *when the data is made of 100 observations* (left column), *1000 observations* (central column) *and 10,000 observations* (right column) *either from* $\mathfrak{M}_1$ (M1) *or* $\mathfrak{M}_2$ (M2) *when the summary statistics in the ABC algorithm are made of the empirical quantile at level* 10% (first row), *the empirical quantiles at levels* 10% *and* 90% (second row), *and the empirical quantiles at levels* 10%, 40%, 60% *and* 90% (third row), *respectively. The boxplots rely on* 100 *replicas and the ABC algorithms are based on* 10^4 *proposals from the prior, with the tolerance being chosen as the* 1% *quantile on the distances.*

Conclusions

- Model selection feasible with ABC
- Choice of summary statistics is paramount
- At best, ABC $\to \pi(. \mid \boldsymbol{T}(\mathbf{y}))$ which concentrates around μ_0

Conclusions

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