Component-wise ABC

ABC with Gibbs steps

Grégoire Clarté*,Christian Robert*†, Robin Ryder*, Julien Stoehr*

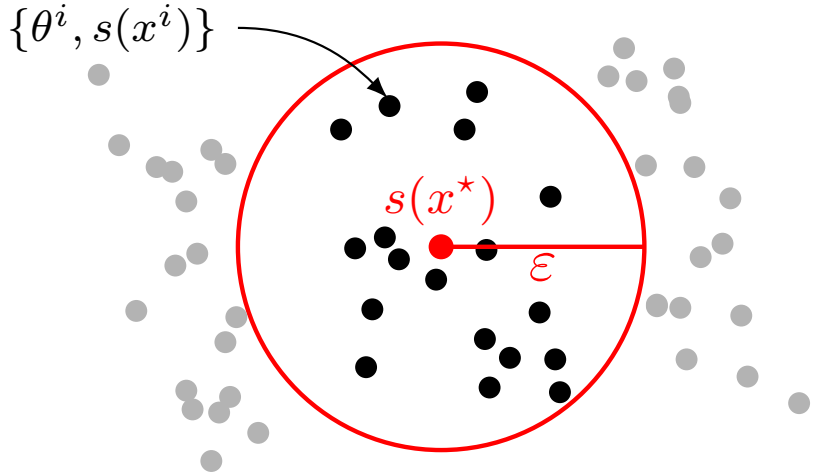
*Université Paris Dauphine ; †Warwick University.

11 juin 2020

ABC

Observations : $x^\star$;
parameters : $\theta_1, \ldots, \theta_n$;
Intractable likelihood
$f(x^\star|\theta_1, \ldots, \theta_n)$.
Possible solution : ABC

- sample $(\theta_1, \ldots, \theta_n)$ from the prior ;
- sample x, pseudo observation, from the likelihood $f(x|\theta_1, \ldots, \theta_n)$;
- keep if $d(s(x), s(x^\star)) < \varepsilon$.

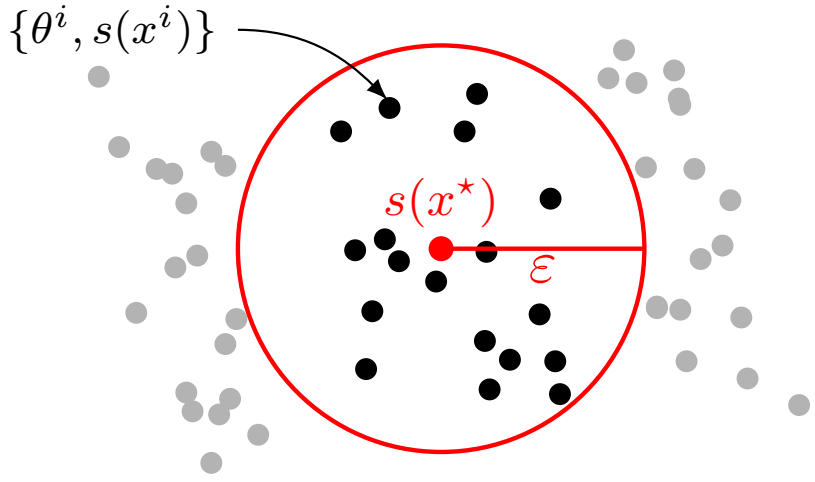


$$\pi_\varepsilon(\theta \mid s, x^\star) \propto \int \pi(\theta) f(x \mid \theta) \mathbf{1}_{d(s(x), s(x^\star)) < \varepsilon} \, \mathrm{d}x$$

ABC

Observations : $x^\star$;
parameters : $\theta_1, \ldots, \theta_n$;
Intractable likelihood
$f(x^\star|\theta_1, \ldots, \theta_n)$.
Possible solution : ABC

- sample $(\theta_1, \ldots, \theta_n)$ from the prior ;
- sample x, pseudo observation, from the likelihood $f(x|\theta_1, \ldots, \theta_n)$;
- keep if $d(s(x), s(x^\star)) < \varepsilon$.

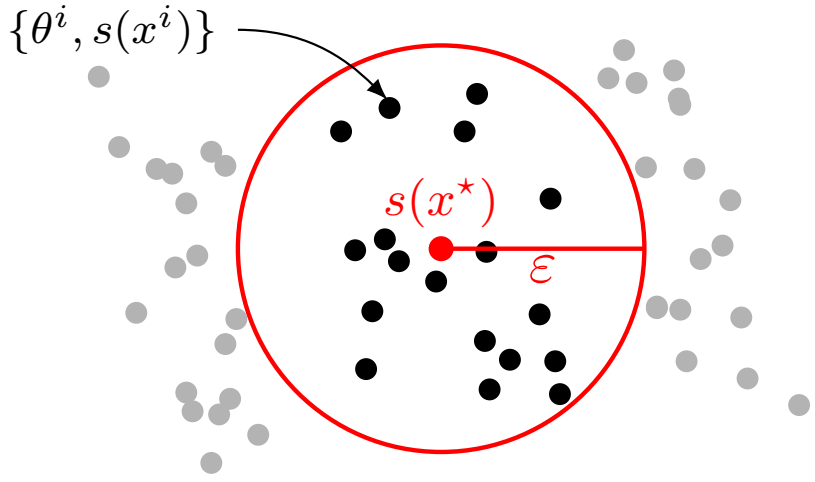


$$\pi_\infty(\theta \mid s, x^\star) \propto \pi(\theta)$$

ABC

Observations : $x^\star$;
parameters : $\theta_1, \ldots, \theta_n$;
Intractable likelihood
$f(x^\star|\theta_1, \ldots, \theta_n)$.
Possible solution : ABC

- sample $(\theta_1, \ldots, \theta_n)$ from the prior ;
- sample x, pseudo observation, from the likelihood $f(x|\theta_1, \ldots, \theta_n)$;
- keep if $d(s(x), s(x^\star)) < \varepsilon$.



$$\pi_0(\theta \mid s, x^\star) \propto \pi(\theta \mid s(x^\star)) \neq \pi(\theta \mid x^\star)$$

Difficulties using ABC

- "Exploration" of parameter space highly inefficient ;
- choice of the summary statistic s, ideally s has same dimension as θ.

Some solutions :

- more complex algorithm, to improve the quality of the proposals (MCMC-ABC) ;
- ABC-Random Forests for the choice of s (only for scalar parameters).

ABC Gibbs

Model described by $\theta = (\theta_1, \ldots, \theta_n)$.

Input: starting point $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)})$, observations x^*.
Output: a sample $(\theta^{(1)}, \ldots, \theta^{(N)})$.
for $i = 1, \ldots, N$ **do**
  for $j = 1, \ldots, n$ **do**
    $\theta_j^{(i)} \sim \pi_{\varepsilon_j}(\cdot \mid x^\star, s_j, \theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, \ldots, \theta_n^{(i-1)})$

Algorithm 1: ABC-Gibbs.

- One tolerance for each parameter ε_j ;
- one statistic for each parameter s_j.

In practice

To run ABC-Gibbs we need to sample one point θ from a law of the form

$$\pi_{\varepsilon_\theta}(\cdot \mid \alpha, s_\theta, x^\star)$$

In practice we use the following procedure:
sample $\theta_1, \dots, \theta_N \sim \pi(\cdot \mid \alpha)$
sample $x_j \sim f(\cdot \mid \theta_j, \alpha)$
compute $d_j = d(s_\theta(x_j, \alpha), s_\theta(x^\star, \alpha))$
return $\theta_{\text{argmin}_j d_j}$
That is we use a quantile of distance instead of ε_θ.

Does it converge (in distribution) ?

Theorem 1

Assume that for all $\ell \leq n$, there exists some $0 < \kappa_\ell < 1/2$, such that

$$\kappa_\ell = \sup_{\boldsymbol{\theta}_{>\ell}, \tilde{\boldsymbol{\theta}}_{>\ell}} \sup_{\boldsymbol{\theta}_{<\ell}} \|\pi_{\varepsilon_\ell}(\cdot \mid x^\star, s_\ell, \boldsymbol{\theta}_{<\ell}, \boldsymbol{\theta}_{>\ell}) - \pi_{\varepsilon_\ell}(\cdot \mid x^\star, s_\ell, \boldsymbol{\theta}_{<\ell}, \tilde{\boldsymbol{\theta}}_{>\ell})\|_{TV}$$

with $\boldsymbol{\theta}_{>\ell} = (\theta_{\ell+1}, \theta_{\ell+2}, \ldots, \theta_n)$, and $\boldsymbol{\theta}_{<\ell} = (\theta_1, \theta_2, \ldots, \theta_{\ell-1})$. Then, the Markov chain produced by ABCG converges geometrically in total variation distance to a stationary distribution ν_ε, with geometric rate $1 - \prod_\ell 2\kappa_\ell$.

Remarks, in practice

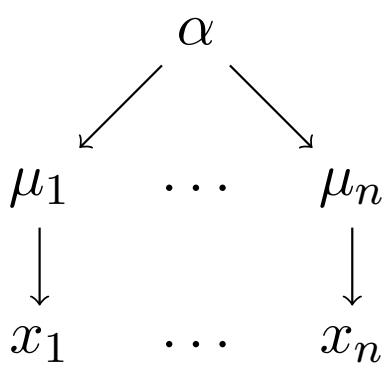
We use stronger assumptions to roughly bound the distances.

- the parameters are compactly supported ;
- the conditional densities never vanish outside of the above mentioned support ;
- the conditional likelihoods are continuous in the parameters.

Furthermore, we affirm that :

- the speed is highly suboptimal ;
- we can prove more particular results for each model.

ABCG hierarchical case



Input: observations $x^\star$, initial value $(\alpha^{(0)}, \mu^{(0)})$

Output: A sample $(\alpha^{(i)}, \mu_1^{(i)}, \ldots, \mu_n^{(i)})_i$.

for $i = 1, \ldots, N$ **do**

  for $j = 1, \ldots, n$ **do**

    $\mu_j^{(i)} \sim \pi_{\varepsilon_{\mu_j}}(\cdot \mid \alpha^{(i-1)}, s_{\mu_j}, x^\star)$

  $\alpha^{(i)} \sim \pi_{\varepsilon_\alpha}(\cdot \mid s_\alpha, \mu_1^{(i)}, \ldots, \mu_n^{(i)})$

Algorithm 2: ABC-Gibbs sampler for hierarchical models.

A toy example

Normal hierarchical model with known variances, 20 parameters, comparison with the exact posterior.

- hyperparameter α with prior $\mathcal{U}([-4, 4])$;
- parameters $\mu_1, \ldots, \mu_{20}$ iid $\mathcal{N}(\alpha, 1)$;
- observations $x_1, \ldots, x_{20}$ iid $\mathcal{N}(\mu_i, 1)^{\otimes 10}$.

ABC Vanilla / ABC-SMC :

- We need *one* summary statistic $s = (\bar{x_1}, \ldots, \bar{x_{20}})$ (sufficient statistic here, the best possible), $d = |\cdot|_1$;
- These methods imply to sample in R^{21}, with a high correlation.
  - the correlation is included in the prior, so simple for ABC vanilla
  - far more difficult with ABC-MCMC / ABC-SMC methods : the proposal kernel must be adapted, even in this example no clue on the optimal kernel.

A toy example : how to Gibbs in practice

In order to run ABC-Gibbs we need to find :

- A "sumary statistic" for each μ_i given μ_{-i}, α, x, where x is a pseudo observation from $f(\cdot \mid \alpha, \mu)$.
- a "sumary statistic" for α given μ, x.
- distances in the space of the statistic

These are not statistic in the classical sense, as they depend on the value of the parameter upon which we condition at each step.
Here, the choice is simple, thanks to the hierarchical structure:

- for all parameters, $s_{\mu_i}(x, \mu_{-i}, \alpha) = \bar{x_i}$;
- for the hyperparameter, $s_\alpha(\mu, x) = \bar{\mu}$;
- any euclidean distance is ok in R.

Why it's a good idea

For the selection of the statistics

- it is usually simpler to find a summary statistic for 1 parameter;
- hopefully the statistic is 1 dimensionned, so no need to find a distance in a strange space;

For the computations

- drastic reduction of parameter dimension;
- often the statistic can be simulated at smaller cost (e.g. hierarchical mode, we only need to simulate "downstream");
- for a given computational time N, we reach a lower tolerance $1/N$ quantile, than vanilla ABC.

A toy example : the algorithm fully developed

Input: observations $x^\star$, initial value $(\alpha^{(0)}, \mu^{(0)})$
Output: A sample $(\alpha^{(i)}, \mu_1^{(i)}, \ldots, \mu_n^{(i)})_i$.
for $i = 1, \ldots, N$ **do**
  for $j = 1, \ldots, n$ **do**
    sample $\mu_j^{1,\ldots,N_\mu} \sim \pi(\cdot \mid \alpha^{(i-1)})$ and associated
     pseudo-observations $x_j^{1,\ldots,N_\mu} \sim f(\cdot \mid \mu_j^{1,\ldots,N_\mu})$.
    $\mu_j^{(i)} \leftarrow \text{argmin}_{k=1,\ldots,N_\mu} d(s_\mu(x_j^k), s_\mu(x_j^\star))$
  sample $\alpha^{1,\ldots,N_\alpha} \sim \pi$ and associated "pseudo-observations"
   $\mu^{1,\ldots,N_\alpha} \sim \pi(\cdot \mid \alpha)$.
  $\alpha^{(i)} \leftarrow \text{argmin}_{k=1,\ldots,N_\alpha} d(s_\alpha(\mu^k), s_\alpha(\mu^{(i)})$

Algorithm 3: ABC-Gibbs sampler for hierarchical models.

Verifying the assumptions

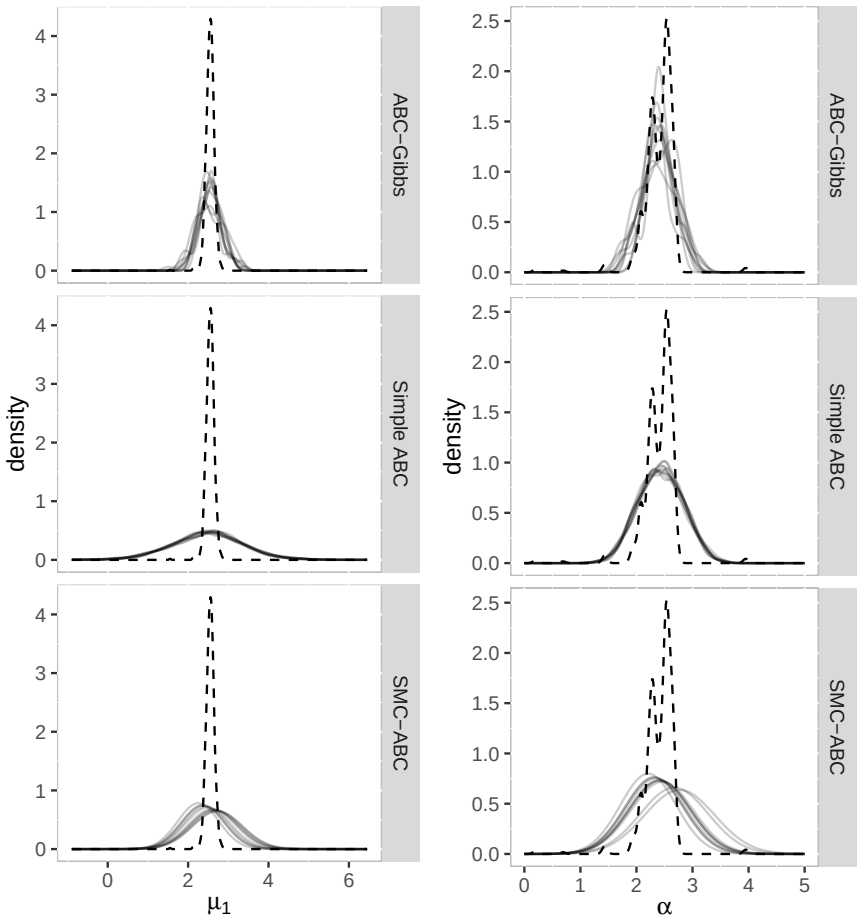
It is sufficient to check : $\exists C$ compact with $\pi_{\varepsilon_\mu}(\mu \mid \alpha, x^\star) > c^{te} > 0, \forall \mu \in C, \forall \alpha$. Here,

$$\pi_\varepsilon(\mu \mid \alpha, s(x^*))$$
$$= \frac{\exp(-(\mu-\alpha)^2/(2\tau) \int \exp(-(y-\mu)^2\sqrt{n}/(2\sigma)\mathbf{1}_{|y-\bar{x^\star}|<\varepsilon} \mathrm{d}y}{\int \exp(-(\mu-\alpha)^2/(2\tau) \exp(-(y-\mu)^2\sqrt{n}/(2\sigma)\mathbf{1}_{|y-\bar{x^\star}|<\varepsilon} \mathrm{d}y \mathrm{d}\mu}$$

as alpha is compactly supported on $[-4, 4]$, the conditions are verified for any compact C: we can roughly bound the probabilities by continuity of the expression.
The last condition on α is always verified as we have by definition of the total variation distance:

$$\sup_\mu \|\pi_{\varepsilon_\alpha}(\cdot \mid \mu) - \pi_{\varepsilon_\alpha}(\cdot \mid \mu)\|_{TV} = 0.$$

Results



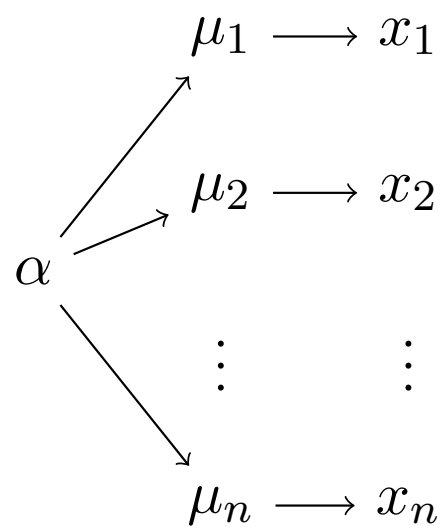
- for ABCG : $N_\mu = N_\alpha = 30$, 10^3 iterations ;
- for ABC vanilla : 1000 points, the best among $3 \cdot 10^4$;
- ABC-SMC with 1000 particles, version adaptive Del Moral, $M = 30$, 500 steps.
  ABCG and ABC vanilla have same computational cost. SMC cost more than 300 times more.

Simulations : Hierarchical G& K

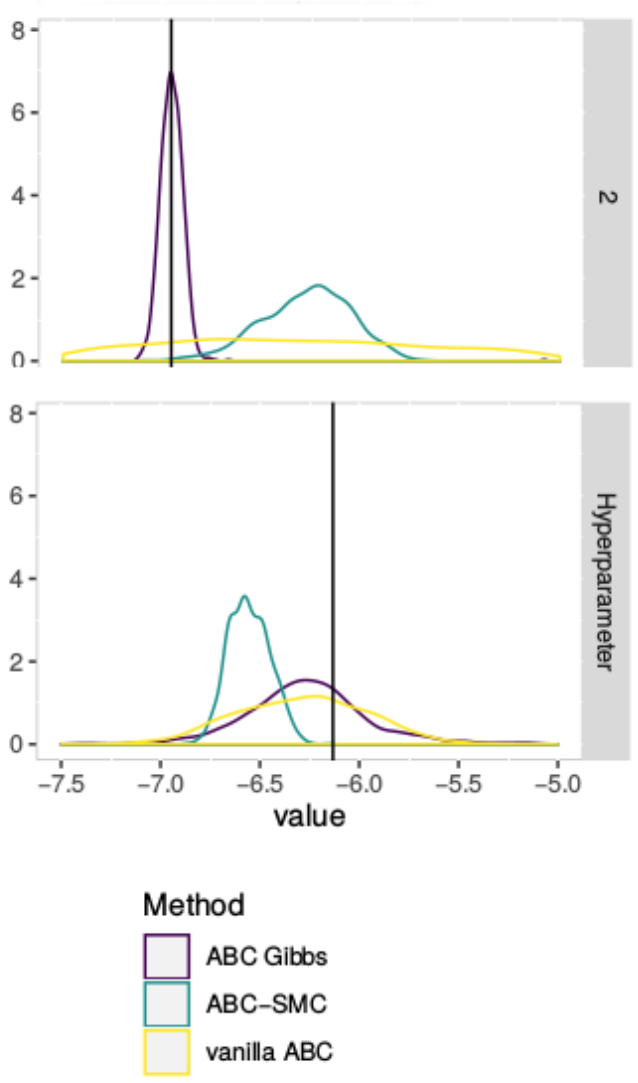
$G\&K$ distribution defined by its quantiles :

$$Q_{gk}(z; \mu, B, g, k, c) = \mu + B(1 + c \tanh[gz/2])z(1 + z^2)^k$$

- $\alpha \in \mathbf{R}$, with prior $\mathcal{U}([-10, 10])$;
- for each j, $\mu_j \sim \mathcal{N}(\alpha, 1)$;
- the other parameters : $B,\ g,\ k$ are known and common to each x_j ;
- in our examples, we have $n = 50$
- Here, the statistics and distances are :
  - $s_\alpha(\mu, x) = \bar{\mu}$
  - $s_{\mu_j}(\mu_{-j}, x, \alpha) = \text{octiles}(x)$, with $d = |\cdot|_1$

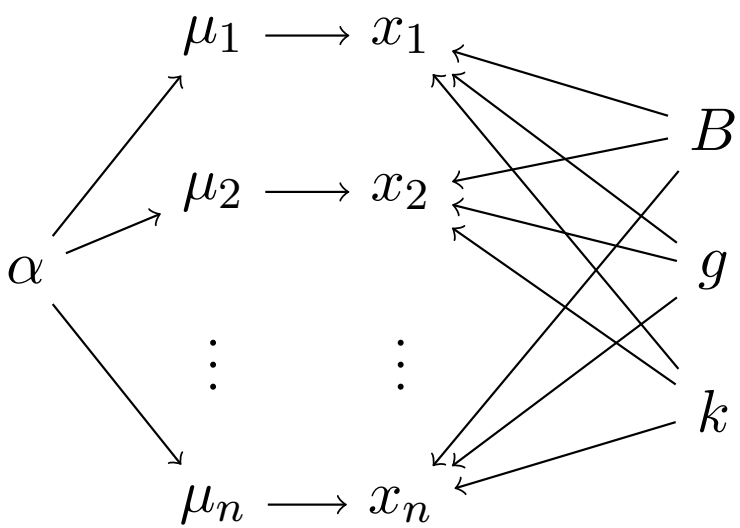


Results



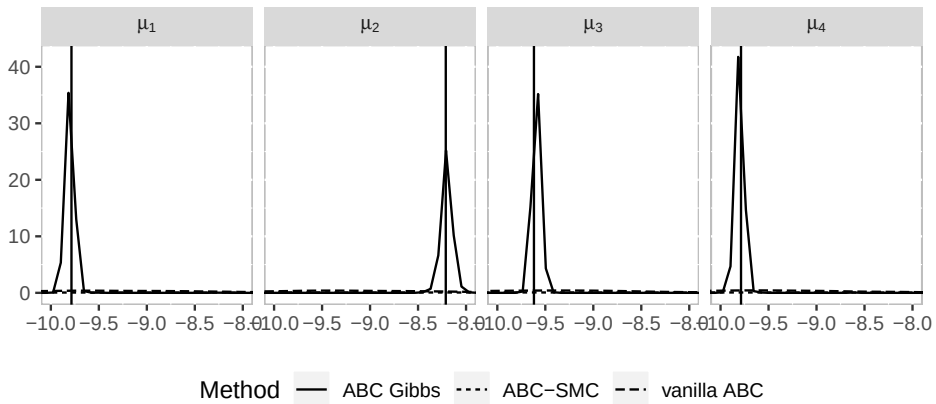
- For a similar computational cost, better results compared to ABC ;
- Simple SMC-ABC fails, parameter too difficult to tune, the particle system degenerates ;
- however this model is not very interesting

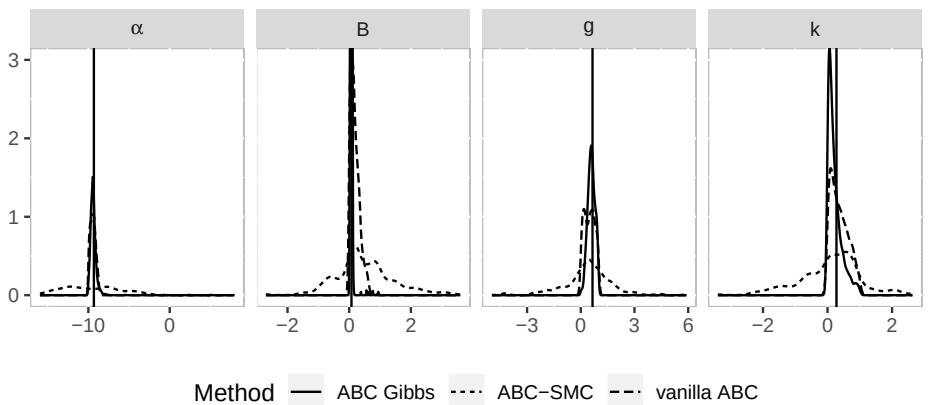
"Hierarchical" G & K



- For μ and α same statistics as before;
- For B, g, k octiles ;
- Comparison with ABC-SMC Del Moral and ABC vanilla, with same computational cost :
  - ABCG : $N_\mu = 100$, $N_\alpha = N_B = N_g = N_k = 50$, 1000 steps ;
  - ABC vanilla : best 1000 points among 10^5 ;
  - ABC-SMC : 1000 particles, $M = 5$, 20 iterations.

Results





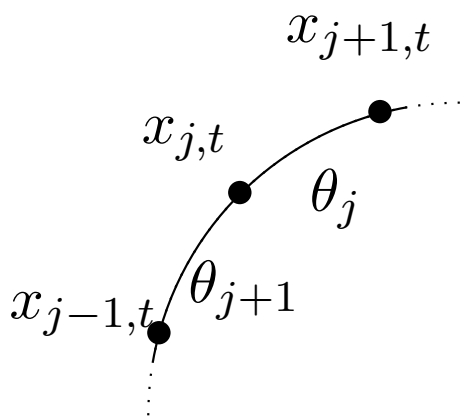
ABC-Gibbs is fine. ABC vanilla returns the prior for the A and a vague posterior for the other parameters. Same for ABC-SMC when it does not degenerates.

Full dependency

Heat equation : $\partial_\tau y(z,\tau) = \partial_z\left(\theta(z)\partial_z y(z,\tau)\right)$.
After discretization : recurring sequence $(y_{j,t})$, with parameter : $\theta = (\theta_1, \ldots, \theta_n)$:

$$\frac{y_{j,t+1} - y_{j,t}}{3\Delta} + \frac{y_{j+1,t+1} - y_{j+1,t}}{6\Delta} + \frac{y_{j-1,t+1} - y_{j-1,t}}{6\Delta}$$
$$= y_{j,t+1}(\theta_{j+1} + \theta_j) - y_{j-1,t+1}\theta_j - y_{j+1,t+1}\theta_{j+1}.$$

Observations : $x_{j,t} = \mathcal{N}(y_{j,t}, \sigma^2)$.
Parameters : θ_j.



Full dependency : how to Gibbs

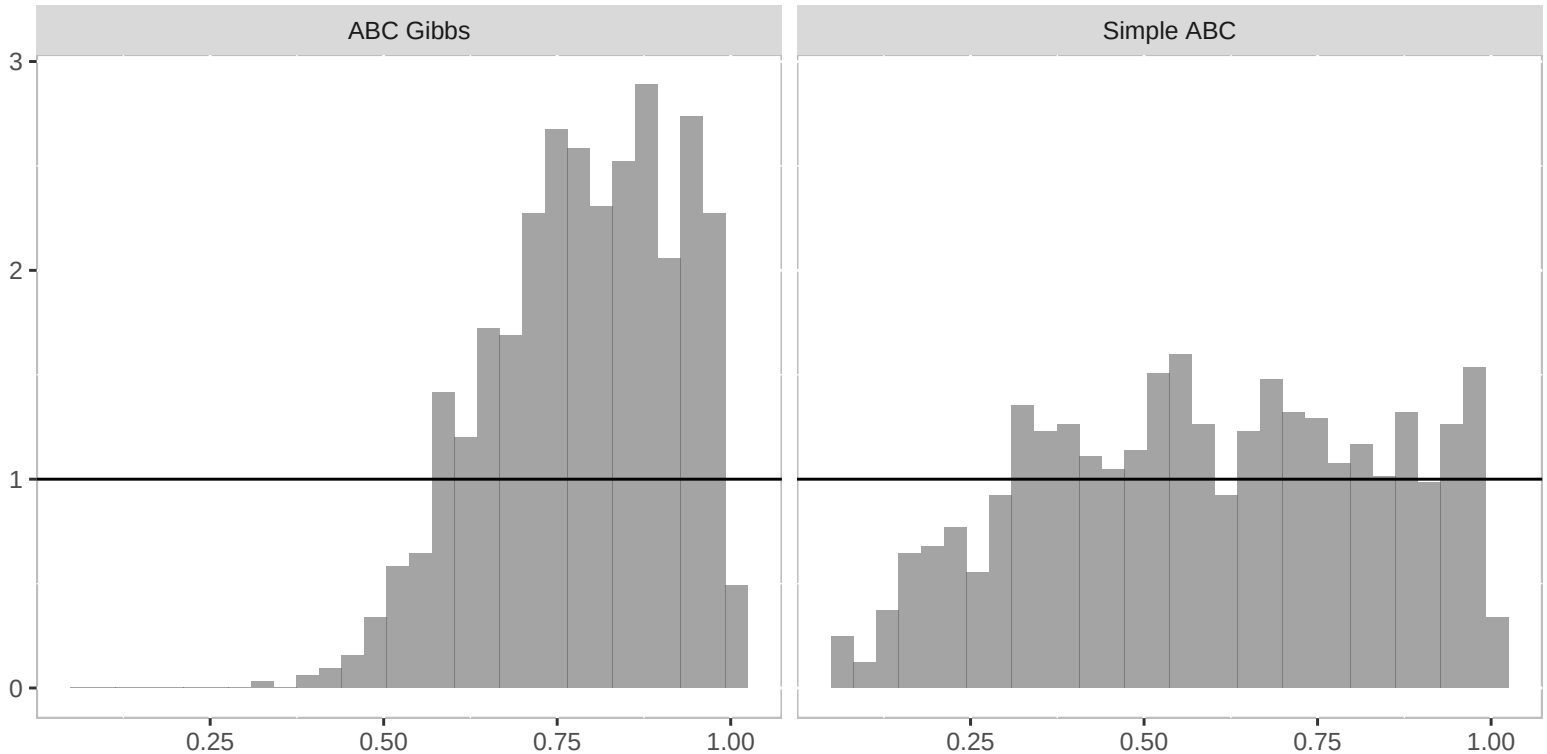
Here, no hierarchical structure $\Rightarrow$ we cannot expect to reduce the size of the simulations.

- in vanilla ABC, $s = Id$, $d = |\cdot|_1$;
- in ABC Gibbs, $s_j(x) = (x_{j-2}, x_{j-1}, x_j, x_{j+1})$, $d = |\cdot|_1$.

As θ_j has a "local" effect, we restrict the statistics to a part of the observations.
$\Rightarrow$ Still smaller dimension.

Results for full dependency model

True value : 0.75, $n = 20$, $N = 8 \cdot 10^6$.



To what does it converge ? hierarchical case, n=2

Let ν_ε be the limiting law of our algorithm, and ν_0 the limiting law of our algorithm for $\varepsilon_\alpha = \varepsilon_\mu = 0$.

Theorem 2 (C *et al.* (2019))

Assume that,

$$L_0 = \sup_{\varepsilon_\alpha} \sup_{\mu,\tilde{\mu}} \|\pi_{\varepsilon_\alpha}(\cdot \mid s_\alpha, \mu) - \pi_0(\cdot \mid s_\alpha, \tilde{\mu})\|_{TV} < 1/2\,,$$

$$L_1(\varepsilon_\mu) = \sup_{\alpha} \|\pi_{\varepsilon_\mu}(\cdot \mid x^\star, s_\mu, \alpha) - \pi_0(\cdot \mid x^\star, s_\mu, \alpha)\|_{TV} \xrightarrow[\varepsilon_\mu \to 0]{} 0\,,$$

$$L_2(\varepsilon_\alpha) = \sup_{\mu} \|\pi_{\varepsilon_\alpha}(\cdot \mid s_\alpha, \mu) - \pi_0(\cdot \mid s_\alpha, \mu)\|_{TV} \xrightarrow[\varepsilon_\alpha \to 0]{} 0\ .$$

Then,

$$\|\nu_\varepsilon - \nu_0\|_{TV} \leq \frac{L_1(\varepsilon_\mu) + L_2(\varepsilon_\alpha)}{1 - 2L_0} \xrightarrow[\varepsilon \to 0]{} 0.$$

Compatibility

ν_0 limiting distribution associated with a Gibbs of conditionals :

$$\pi(\alpha)\pi(s_\alpha(\mu) \mid \alpha) \text{ and } \pi(\mu)f(s_\mu(x^\star) \mid \alpha, \mu).$$

They can be incompatible.
If s_α is sufficient, when $\varepsilon_\alpha \to 0$ the limiting distribution is the same as ABC.

Open questions

- What can be said about the incompatible case ?
  $\Rightarrow$ it seems that the prior constrains the approximate posterior to be a true density.
- choose s_α et s_μ ? $\Rightarrow$ small dimensioned and "locally" informative (*i.e.* conditionally to the value of the parameters) ;
- weaken the assumptions of the theorems ;
- adapt the result to other approximations of the conditionals.

Bibliography & code

Component-wise approximate Bayesian computation via Gibbs-like steps, Grégoire Clarté, Christian P. Robert, Robin Ryder, Julien Stoehr.
`arXiv:1905.13599`

`https://github.com/GClarte/ABCG`