Part 3: Some recent advances on Approximate Bayesian Computation

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck (IMAG)

September, 22-27, Fréjus, France

Thanks

Numerous colleagues participated to parts of this work

- Pierre Pudlo (Marseille)
- Louis Raynal (PhD student Montpellier, postdoc Harvard)
- Arnaud Estoup (molecular ecologist, Montpellier)
- Christian Robert (Paris and Warwick)
- Judith, Natesh, ...

Thanks

Numerous colleagues participated to parts of this work

- Pierre Pudlo (Marseille)
- Louis Raynal (PhD student Montpellier, postdoc Harvard)
- Arnaud Estoup (molecular ecologist, Montpellier)
- Christian Robert (Paris and Warwick)
- Judith, Natesh, ...

Thanks

Numerous colleagues participated to parts of this work

- Pierre Pudlo (Marseille)
- Louis Raynal (PhD student Montpellier, postdoc Harvard)
- Arnaud Estoup (molecular ecologist, Montpellier)
- Christian Robert (Paris and Warwick)
- Judith, Natesh, ...

Thanks

Numerous colleagues participated to parts of this work

- Pierre Pudlo (Marseille)
- Louis Raynal (PhD student Montpellier, postdoc Harvard)
- Arnaud Estoup (molecular ecologist, Montpellier)
- Christian Robert (Paris and Warwick)
- Judith, Natesh, ...

Thanks

Numerous colleagues participated to parts of this work

- Pierre Pudlo (Marseille)
- Louis Raynal (PhD student Montpellier, postdoc Harvard)
- Arnaud Estoup (molecular ecologist, Montpellier)
- Christian Robert (Paris and Warwick)
- Judith, Natesh, ...

Thanks

Numerous colleagues participated to parts of this work

- Pierre Pudlo (Marseille)
- Louis Raynal (PhD student Montpellier, postdoc Harvard)
- Arnaud Estoup (molecular ecologist, Montpellier)
- Christian Robert (Paris and Warwick)
- Judith, Natesh, ...

Introduction

Bayesian parametric paradigm

Likelihood function $f(\mathbf{y}|\theta)$ expensive or impossible to calculate

Extremely difficult to sample from the posterior distribution

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

Introduction

Bayesian parametric paradigm

Likelihood function $f(\mathbf{y}|\theta)$ expensive or impossible to calculate

Extremely difficult to sample from the posterior distribution

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

Introduction

Bayesian parametric paradigm

Likelihood function $f(\mathbf{y}|\theta)$ expensive or impossible to calculate

Extremely difficult to sample from the posterior distribution

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

Introduction

Two typical situations

- $f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{u}|\theta)\mu(d\mathbf{u})$ intractable
  population genetics models, coalescent process

  EM algorithms, Gibbs sampling, pseudo-marginal MCMC methods, variational approximations

- $f(\mathbf{y}|\theta) = g(\mathbf{y}, \theta)/Z(\theta)$ and $Z(\theta)$ intractable
  Markov random field

  pseudo-marginal MCMC methods, variational approximations

Introduction

Two typical situations

- $f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{u}|\theta)\mu(d\mathbf{u})$ intractable

  population genetics models, coalescent process

  EM algorithms, Gibbs sampling, pseudo-marginal MCMC methods, variational approximations

- $f(\mathbf{y}|\theta) = g(\mathbf{y}, \theta)/Z(\theta)$ and $Z(\theta)$ intractable

  Markov random field

  pseudo-marginal MCMC methods, variational approximations

Introduction

Two typical situations

- $f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{u}|\theta)\mu(d\mathbf{u})$ intractable
  population genetics models, coalescent process

  EM algorithms, Gibbs sampling, pseudo-marginal MCMC methods, variational approximations

- $f(\mathbf{y}|\theta) = g(\mathbf{y}, \theta)/Z(\theta)$ and $Z(\theta)$ intractable
  Markov random field

  pseudo-marginal MCMC methods, variational approximations

Introduction

Two typical situations

- $f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{u}|\theta)\mu(d\mathbf{u})$ intractable
  population genetics models, coalescent process

  EM algorithms, Gibbs sampling, pseudo-marginal MCMC methods, variational approximations

- $f(\mathbf{y}|\theta) = g(\mathbf{y}, \theta)/Z(\theta)$ and $Z(\theta)$ intractable
  Markov random field

  pseudo-marginal MCMC methods, variational approximations

Introduction

Two typical situations

- $f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{u}|\theta)\mu(d\mathbf{u})$ intractable
  population genetics models, coalescent process

  EM algorithms, Gibbs sampling, pseudo-marginal MCMC methods, variational approximations

- $f(\mathbf{y}|\theta) = g(\mathbf{y}, \theta)/Z(\theta)$ and $Z(\theta)$ intractable
  Markov random field

  pseudo-marginal MCMC methods, variational approximations

Introduction

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

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

If, with Christian, we work on ABC methods, we can be very grateful to our biologist colleagues!

Introduction

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

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

If, with Christian, we work on ABC methods, we can be very grateful to our biologist colleagues!

Introduction

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

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

If, with Christian, we work on ABC methods, we can be very grateful to our biologist colleagues!

Introduction

- some methodological aspects of ABC
- our ABC random forests proposal
- a population genetics example

Introduction

- some methodological aspects of ABC
- our ABC random forests proposal
- a population genetics example

Introduction

- some methodological aspects of ABC
- our ABC random forests proposal
- a population genetics example

1. Methodological aspects of ABC
   - Likelihood-free rejection sampler
   - A k-NN approximation
   - Two views of the ABC approximation
   - More efficient algorithms
   - Regression adjustments
   - Summary statistics
   - ABC model choice procedure
   - Sofwares
   - Frontline news from population geneticists country
   - Use modern machine learning tools

2. ABC random forests
   - Model choice
   - Parameter inference

3. Human populations example

Methodological aspects of ABC
Likelihood-free rejection sampler

Rubin (1984) The Annals of Statistics
Tavaré et al. (1997) Genetics
Pritchard et al. (1999) Mol. Biol. Evol.

```
1) Set $i = 1$
2) Generate $\theta'$ from the prior distribution $\pi(\cdot)$
3) Generate $\mathbf{z}$ from the likelihood $f(\cdot|\theta')$
4) If $d(\eta(\mathbf{z}), \eta(\mathbf{y})) \leqslant \epsilon$, set $\theta_i = \theta'$ and $i = i + 1$
5) If $i \leqslant N$, return to 2)
```

Methodological aspects of ABC
Likelihood-free rejection sampler

Rubin (1984) The Annals of Statistics
Tavaré et al. (1997) Genetics
Pritchard et al. (1999) Mol. Biol. Evol.

```
1) Set $i = 1$
2) Generate $\theta'$ from the prior distribution $\pi(\cdot)$
3) Generate $z$ from the likelihood $f(\cdot|\theta')$
4) If $d(\eta(z), \eta(y)) \leqslant \epsilon$, set $\theta_i = \theta'$ and $i = i + 1$
5) If $i \leqslant N$, return to 2)
```

Methodological aspects of ABC
Likelihood-free rejection sampler

Rubin (1984) The Annals of Statistics
Tavaré et al. (1997) Genetics
Pritchard et al. (1999) Mol. Biol. Evol.

1) Set $i = 1$
2) Generate θ' from the prior distribution $\pi(\cdot)$
3) Generate z from the likelihood $f(\cdot|\theta')$
4) If $d(\eta(\mathbf{z}), \eta(\mathbf{y})) \leqslant \epsilon$, set $\theta_i = \theta'$ and $i = i + 1$
5) If $i \leqslant N$, return to **2)**

Methodological aspects of ABC
Likelihood-free rejection sampler

Rubin (1984) The Annals of Statistics
Tavaré et al. (1997) Genetics
Pritchard et al. (1999) Mol. Biol. Evol.

```
1) Set $i = 1$
2) Generate $\theta'$ from the prior distribution $\pi(\cdot)$
3) Generate $\mathbf{z}$ from the likelihood $f(\cdot|\theta')$
4) If $d(\eta(\mathbf{z}), \eta(\mathbf{y})) \leqslant \epsilon$, set $\theta_i = \theta'$ and $i = i + 1$
5) If $i \leqslant N$, return to 2)
```

Methodological aspects of ABC
Likelihood-free rejection sampler

Rubin (1984) The Annals of Statistics
Tavaré et al. (1997) Genetics
Pritchard et al. (1999) Mol. Biol. Evol.

```
1) Set $i = 1$
2) Generate $\theta'$ from the prior distribution $\pi(\cdot)$
3) Generate $\mathbf{z}$ from the likelihood $f(\cdot|\theta')$
4) If $d(\eta(\mathbf{z}), \eta(\mathbf{y})) \leqslant \epsilon$, set $\theta_i = \theta'$ and $i = i + 1$
5) If $i \leqslant N$, return to 2)
```

Methodological aspects of ABC
Likelihood-free rejection sampler

Rubin (1984) The Annals of Statistics
Tavaré et al. (1997) Genetics
Pritchard et al. (1999) Mol. Biol. Evol.

1) Set $i = 1$

2) Generate θ' from the prior distribution $\pi(\cdot)$

3) Generate $\mathbf{z}$ from the likelihood $f(\cdot|\theta')$

4) If $d(\eta(\mathbf{z}), \eta(\mathbf{y})) \leqslant \epsilon$, set $\theta_i = \theta'$ and $i = i + 1$

5) If $i \leqslant N$, return to **2)**

Methodological aspects of ABC
Likelihood-free rejection sampler

ϵ reflects the tension between computability and accuracy

- if $\epsilon \to \infty$, we get simulations from the prior
- if $\epsilon \to 0$, we get simulations from the posterior

ABC target

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

$A_{\epsilon,\mathbf{y}} = \{\mathbf{z} | d(\eta(\mathbf{z}), \eta(\mathbf{y})) \leqslant \epsilon\}$ the acceptance set

Methodological aspects of ABC
Likelihood-free rejection sampler

ϵ reflects the tension between computability and accuracy

- if $\epsilon \to \infty$, we get simulations from the prior
- if $\epsilon \to 0$, we get simulations from the posterior

ABC target

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

$A_{\epsilon,\mathbf{y}} = \{\mathbf{z}|d(\eta(\mathbf{z}),\eta(\mathbf{y})) \leqslant \epsilon\}$ the acceptance set

Methodological aspects of ABC
Likelihood-free rejection sampler

ϵ reflects the tension between computability and accuracy

- if $\epsilon \to \infty$, we get simulations from the prior
- if $\epsilon \to 0$, we get simulations from the posterior

ABC target

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

$A_{\epsilon,\mathbf{y}} = \{\mathbf{z}|d(\eta(\mathbf{z}),\eta(\mathbf{y})) \leqslant \epsilon\}$ the acceptance set

Methodological aspects of ABC
Likelihood-free rejection sampler

ϵ reflects the tension between computability and accuracy

- if $\epsilon \to \infty$, we get simulations from the prior
- if $\epsilon \to 0$, we get simulations from the posterior

ABC target

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

$A_{\epsilon,\mathbf{y}} = \{\mathbf{z} | d(\eta(\mathbf{z}), \eta(\mathbf{y})) \leqslant \epsilon\}$ the acceptance set

Methodological aspects of ABC
Likelihood-free rejection sampler

A toy example from Richard Wilkinson (Tutorial on ABC, NIPS 2013)

$$y|\theta \sim \mathcal{N}_1\left(2(\theta+2)\theta(\theta-2), 0.1+\theta^2\right)$$

$$\theta \sim \mathcal{U}_{[-10,10]}$$

$$y = 2$$

$$d(z,y) = |z-y|$$

Methodological aspects of ABC
Likelihood-free rejection sampler

A toy example from Richard Wilkinson (Tutorial on ABC, NIPS 2013)

$$y|\theta \sim \mathcal{N}_1\left(2(\theta+2)\theta(\theta-2), 0.1+\theta^2\right)$$

$$\theta \sim \mathcal{U}_{[-10,10]}$$

$$y = 2$$

$$d(z,y) = |z-y|$$

Methodological aspects of ABC
Likelihood-free rejection sampler

A toy example from Richard Wilkinson (Tutorial on ABC, NIPS 2013)

$$y|\theta \sim \mathscr{N}_1\left(2(\theta+2)\theta(\theta-2), 0.1+\theta^2\right)$$

$$\theta \sim \mathscr{U}_{[-10,10]}$$

$$y = 2$$

$$d(z,y) = |z-y|$$

Methodological aspects of ABC
Likelihood-free rejection sampler

A toy example from Richard Wilkinson (Tutorial on ABC, NIPS 2013)

$$y|\theta \sim \mathcal{N}_1\left(2(\theta+2)\theta(\theta-2), 0.1+\theta^2\right)$$

$$\theta \sim \mathcal{U}_{[-10,10]}$$

$$y = 2$$

$$d(z,y) = |z-y|$$

Methodological aspects of ABC
Likelihood-free rejection sampler

A toy example from Richard Wilkinson (Tutorial on ABC, NIPS 2013)

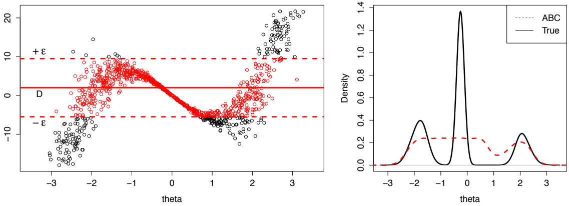
$$y|\theta \sim \mathcal{N}_1\left(2(\theta+2)\theta(\theta-2), 0.1+\theta^2\right)$$

$$\theta \sim \mathcal{U}_{[-10,10]}$$

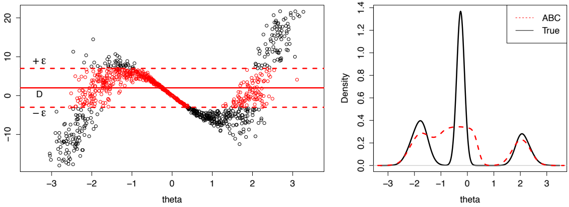
$$y = 2$$

$$d(z, y) = |z - y|$$

Methodological aspects of ABC
Likelihood-free rejection sampler

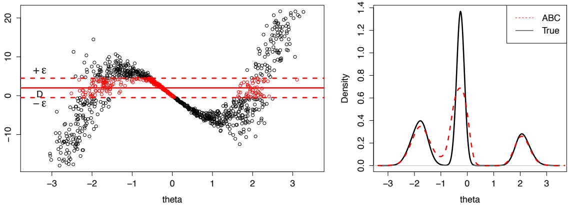


$\epsilon = 7.5$

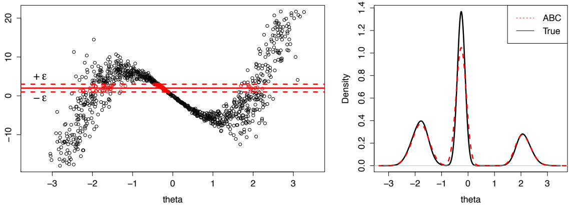


$\epsilon = 5$

Methodological aspects of ABC
Likelihood-free rejection sampler



$\epsilon = 2.5$



$\epsilon = 1$

Methodological aspects of ABC
A k-NN approximation

Practitioners really use

```
1) For $i = 1, \dots, M$
        Generate $\theta_i$ from the prior $\pi(\cdot)$

2) Order the distances $d_{(1)}, \dots, d_{(M)}$
3) Return the $\theta_i$'s that correspond to the N-smallest distances
```

$N = \lfloor \alpha M \rfloor$

ϵ corresponds to a quantile of the distances

Methodological aspects of ABC
A k-NN approximation

Practitioners really use

```
1) For $i = 1, \dots, M$
    a) Generate $\theta_i$ from the prior $\pi(\cdot)$
    b) Generate $\mathbf{z}$ from the model $f(\cdot|\theta_i)$
    c) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \dots, d_{(M)}$
3) Return the $\theta_i$'s that correspond to the N-smallest distances
```

$N = \lfloor \alpha M \rfloor$

ϵ corresponds to a quantile of the distances

Methodological aspects of ABC
A k-NN approximation

Practitioners really use

```
1) For $i = 1, \dots, M$
    a) Generate $\theta_i$ from the prior $\pi(\cdot)$
    b) Generate $\mathbf{z}$ from the model $f(\cdot|\theta_i)$
    c) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \dots, d_{(M)}$
3) Return the $\theta_i$'s that correspond to the N-smallest distances
```

$N = \lfloor \alpha M \rfloor$

ϵ corresponds to a quantile of the distances

Methodological aspects of ABC
A k-NN approximation

Practitioners really use

```
1) For $i = 1, \dots, M$
    a) Generate $\theta_i$ from the prior $\pi(\cdot)$
    b) Generate $\mathbf{z}$ from the model $f(\cdot|\theta_i)$
    c) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \dots, d_{(M)}$
3) Return the $\theta_i$'s that correspond to the N-smallest distances
```

$N = \lfloor \alpha M \rfloor$

ϵ corresponds to a quantile of the distances

Methodological aspects of ABC
A k-NN approximation

Practitioners really use

1) For $i = 1, \dots, M$
   - a) Generate θ_i from the prior $\pi(\cdot)$
   - b) Generate $\mathbf{z}$ from the model $f(\cdot|\theta_i)$
   - c) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \dots, d_{(M)}$
3) Return the θ_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

ϵ corresponds to a quantile of the distances

Methodological aspects of ABC
A k-NN approximation

Practitioners really use

1) For $i = 1, \ldots, M$
   a) Generate θ_i from the prior $\pi(\cdot)$
   b) Generate $\mathbf{z}$ from the model $f(\cdot|\theta_i)$
   c) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$
3) Return the θ_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

ϵ corresponds to a quantile of the distances

Methodological aspects of ABC
A k-NN approximation

Practitioners really use

```
1) For $i = 1, \dots, M$
    a) Generate $\theta_i$ from the prior $\pi(\cdot)$
    b) Generate $\mathbf{z}$ from the model $f(\cdot|\theta_i)$
    c) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \dots, d_{(M)}$
3) Return the $\theta_i$'s that correspond to the $N$-smallest distances
```

$N = \lfloor \alpha M \rfloor$

ϵ corresponds to a quantile of the distances

Methodological aspects of ABC
A k-NN approximation

Practitioners really use

```
1) For $i = 1, \dots, M$
    a) Generate $\theta_i$ from the prior $\pi(\cdot)$
    b) Generate $\mathbf{z}$ from the model $f(\cdot|\theta_i)$
    c) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \dots, d_{(M)}$
3) Return the $\theta_i$'s that correspond to the $N$-smallest distances
```

$N = \lfloor \alpha M \rfloor$

ϵ corresponds to a quantile of the distances

Methodological aspects of ABC
A k-NN approximation

Practitioners really use

1) For $i = 1, \ldots, M$
   a) Generate θ_i from the prior $\pi(\cdot)$
   b) Generate $\mathbf{z}$ from the model $f(\cdot|\theta_i)$
   c) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$
3) Return the θ_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

ϵ corresponds to a quantile of the distances

Methodological aspects of ABC
A k-NN approximation

New insights into Approximate Bayesian Computation
Biau, Cérou, Guyader (2015) Annales de l'IHP

- intuitive
- simple to implement
- embarrassingly parallelisable
- BUT curse of dimensionality: most of the simulations are at the boundary of the space as the number of summary statistics increases

Methodological aspects of ABC
A k-NN approximation

New insights into Approximate Bayesian Computation
Biau, Cérou, Guyader (2015) Annales de l'IHP

- intuitive
- simple to implement
- embarrassingly parallelisable
- BUT curse of dimensionality: most of the simulations are at the boundary of the space as the number of summary statistics increases

Methodological aspects of ABC
A k-NN approximation

New insights into Approximate Bayesian Computation
Biau, Cérou, Guyader (2015) Annales de l'IHP

- intuitive
- simple to implement
- embarrassingly parallelisable
- BUT curse of dimensionality: most of the simulations are at the boundary of the space as the number of summary statistics increases

Methodological aspects of ABC
A k-NN approximation

New insights into Approximate Bayesian Computation
Biau, Cérou, Guyader (2015) Annales de l'IHP

- intuitive
- simple to implement
- embarrassingly parallelisable
- BUT curse of dimensionality: most of the simulations are at the boundary of the space as the number of summary statistics increases

Methodological aspects of ABC
A k-NN approximation

New insights into Approximate Bayesian Computation
Biau, Cérou, Guyader (2015) Annales de l'IHP

- intuitive
- simple to implement
- embarrassingly parallelisable
- BUT curse of dimensionality: most of the simulations are at the boundary of the space as the number of summary statistics increases

New insights into Approximate Bayesian Computation
Biau, Cérou, Guyader (2015) Annales de l'IHP

- intuitive
- simple to implement
- embarrassingly parallelisable
- BUT curse of dimensionality: most of the simulations are at the boundary of the space as the number of summary statistics increases

Methodological aspects of ABC
Two views of the ABC approximation

$\Longrightarrow$ **Wilkinson (2013) SAGMB** shows that ABC is exact but for a different model to that intended

$\Longrightarrow$ **Blum (2010) JASA** emphasizes that ABC is a kernel smoothing approximation of the likelihood function

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

$$= \frac{\pi(\theta) \int f(\mathbf{z}|\theta) K(d(\eta(\mathbf{z}), \eta(\mathbf{y}))) \mathrm{d}\mathbf{z}}{\int \pi(\theta) f(\mathbf{z}|\theta) K(d(\eta(\mathbf{z}), \eta(\mathbf{y}))) \mathrm{d}\mathbf{z} \mathrm{d}\theta}$$

Methodological aspects of ABC
Two views of the ABC approximation

$\Longrightarrow$ **Wilkinson (2013) SAGMB** shows that ABC is exact but for a different model to that intended

$\Longrightarrow$ **Blum (2010) JASA** emphasizes that ABC is a kernel smoothing approximation of the likelihood function

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

$$= \frac{\pi(\theta)\int f(\mathbf{z}|\theta)K(d(\eta(\mathbf{z}),\eta(\mathbf{y})))d\mathbf{z}}{\int \pi(\theta) f(\mathbf{z}|\theta)K(d(\eta(\mathbf{z}),\eta(\mathbf{y})))d\mathbf{z}d\theta}$$

Methodological aspects of ABC
Two views of the ABC approximation

$\Longrightarrow$ **Wilkinson (2013) SAGMB** shows that ABC is exact but for a different model to that intended

$\Longrightarrow$ **Blum (2010) JASA** emphasizes that ABC is a kernel smoothing approximation of the likelihood function

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

$$= \frac{\pi(\theta) \int f(\mathbf{z}|\theta) K(d(\eta(\mathbf{z}), \eta(\mathbf{y}))) \mathrm{d}\mathbf{z}}{\int \pi(\theta) f(\mathbf{z}|\theta) K(d(\eta(\mathbf{z}), \eta(\mathbf{y}))) \mathrm{d}\mathbf{z} \mathrm{d}\theta}$$

Methodological aspects of ABC
More efficient algorithms

Simulate all the θ's particles using the prior distribution

$\Longrightarrow$ very inefficient

various sequential Monte Carlo algorithms have been constructed as an alternative

Sisson et al. (2007) PNAS
Beaumont, Cornuet, Marin and Robert (2009) Biometrika
Del Moral et al. (2012) Statistics and Computing
Marin, Pudlo and Sedki (2012) IEEE Proceedings of WSC
Filippi et al. (2013) SAGMB

Methodological aspects of ABC
More efficient algorithms

Simulate all the θ's particles using the prior distribution

$\Longrightarrow$ very inefficient

various sequential Monte Carlo algorithms have been constructed as an alternative

Sisson et al. (2007) PNAS
Beaumont, Cornuet, Marin and Robert (2009) Biometrika
Del Moral et al. (2012) Statistics and Computing
Marin, Pudlo and Sedki (2012) IEEE Proceedings of WSC
Filippi et al. (2013) SAGMB

Methodological aspects of ABC
More efficient algorithms

Simulate all the θ's particles using the prior distribution

$\Longrightarrow$ very inefficient

various sequential Monte Carlo algorithms have been constructed as an alternative

Sisson et al. (2007) PNAS
Beaumont, Cornuet, Marin and Robert (2009) Biometrika
Del Moral et al. (2012) Statistics and Computing
Marin, Pudlo and Sedki (2012) IEEE Proceedings of WSC
Filippi et al. (2013) SAGMB

Methodological aspects of ABC
More efficient algorithms

Simulate all the θ's particles using the prior distribution

$\Longrightarrow$ very inefficient

various sequential Monte Carlo algorithms have been constructed as an alternative

Sisson et al. (2007) PNAS
Beaumont, Cornuet, Marin and Robert (2009) Biometrika
Del Moral et al. (2012) Statistics and Computing
Marin, Pudlo and Sedki (2012) IEEE Proceedings of WSC
Filippi et al. (2013) SAGMB

Methodological aspects of ABC
More efficient algorithms

Simulate all the θ's particles using the prior distribution

$\Longrightarrow$ very inefficient

various sequential Monte Carlo algorithms have been constructed as an alternative

Sisson et al. (2007) PNAS
Beaumont, Cornuet, Marin and Robert (2009) Biometrika
Del Moral et al. (2012) Statistics and Computing
Marin, Pudlo and Sedki (2012) IEEE Proceedings of WSC
Filippi et al. (2013) SAGMB

Methodological aspects of ABC
More efficient algorithms

The key idea is to decompose the difficult problem of sampling from $\pi_\epsilon(\theta, \mathbf{z}|\mathbf{y})$ into a series of simpler subproblems

Time 0 sampling from $\pi_{\epsilon_0}(\theta, \mathbf{z}|\mathbf{y})$ with large ϵ_0
Then simulating from an increasing difficult sequence of target distribution $\pi_{\epsilon_t}(\theta, \mathbf{z}|\mathbf{y})$ that is $\epsilon_t < \epsilon_{t-1}$

Likelihood free MCMC sampler **Majoram et al. (2003) PNAS**

Methodological aspects of ABC
More efficient algorithms

The key idea is to decompose the difficult problem of sampling from $\pi_\epsilon(\theta, \mathbf{z}|\mathbf{y})$ into a series of simpler subproblems

Time 0 sampling from $\pi_{\epsilon_0}(\theta, \mathbf{z}|\mathbf{y})$ with large ϵ_0
Then simulating from an increasing difficult sequence of target distribution $\pi_{\epsilon_t}(\theta, \mathbf{z}|\mathbf{y})$ that is $\epsilon_t < \epsilon_{t-1}$

Likelihood free MCMC sampler **Majoram et al. (2003) PNAS**

Methodological aspects of ABC
More efficient algorithms

The key idea is to decompose the difficult problem of sampling from $\pi_\epsilon(\theta, \mathbf{z}|\mathbf{y})$ into a series of simpler subproblems

Time 0 sampling from $\pi_{\epsilon_0}(\theta, \mathbf{z}|\mathbf{y})$ with large ϵ_0
Then simulating from an increasing difficult sequence of target distribution $\pi_{\epsilon_t}(\theta, \mathbf{z}|\mathbf{y})$ that is $\epsilon_t < \epsilon_{t-1}$

Likelihood free MCMC sampler **Majoram et al. (2003) PNAS**

Methodological aspects of ABC
More efficient algorithms

The key idea is to decompose the difficult problem of sampling from $\pi_\epsilon(\theta, \mathbf{z}|\mathbf{y})$ into a series of simpler subproblems

Time 0 sampling from $\pi_{\epsilon_0}(\theta, \mathbf{z}|\mathbf{y})$ with large ϵ_0
Then simulating from an increasing difficult sequence of target distribution $\pi_{\epsilon_t}(\theta, \mathbf{z}|\mathbf{y})$ that is $\epsilon_t < \epsilon_{t-1}$

Likelihood free MCMC sampler **Majoram et al. (2003) PNAS**

Methodological aspects of ABC
Regression adjustments

Beaumont et al. (2002) Genetics
local linear regression adjustment of the parameter values

Blum and Francois (2010) Statistics and Computing
heteroscedastic models, feed-forward neural networks

Methodological aspects of ABC
Regression adjustments

Beaumont et al. (2002) Genetics
local linear regression adjustment of the parameter values

Blum and Francois (2010) Statistics and Computing
heteroscedastic models, feed-forward neural networks

Methodological aspects of ABC
Regression adjustments

Beaumont et al. (2002) Genetics
local linear regression adjustment of the parameter values

Blum and Francois (2010) Statistics and Computing
heteroscedastic models, feed-forward neural networks

Methodological aspects of ABC
Summary statistics

Best subset selection

- **Joyce and Marjoram (2008) SAGMB**, τ-sufficiency
- **Nunes and Balding (2010) SAGMB**, entropy

Projection

- **Fearnhead and Prangle (2012) JRSS B** introduce semi-automatic ABC

Regularization techniques

- **Blum, Nunes, Prangle and Fearnhead (2013) Statistical Science** use ridge regression
- **Saulnier, Gascuel, Alizon (2017) Plos Computational Biology** use LASSO

Methodological aspects of ABC
Summary statistics

Best subset selection

- **Joyce and Marjoram (2008) SAGMB**, τ-sufficiency
- **Nunes and Balding (2010) SAGMB**, entropy

Projection

- **Fearnhead and Prangle (2012) JRSS B** introduce semi-automatic ABC

Regularization techniques

- **Blum, Nunes, Prangle and Fearnhead (2013) Statistical Science** use ridge regression
- **Saulnier, Gascuel, Alizon (2017) Plos Computational Biology** use LASSO

Methodological aspects of ABC
Summary statistics

Best subset selection

- **Joyce and Marjoram (2008) SAGMB**, τ-sufficiency
- **Nunes and Balding (2010) SAGMB**, entropy

Projection

- **Fearnhead and Prangle (2012) JRSS B** introduce semi-automatic ABC

Regularization techniques

- **Blum, Nunes, Prangle and Fearnhead (2013) Statistical Science** use ridge regression
- **Saulnier, Gascuel, Alizon (2017) Plos Computational Biology** use LASSO

Methodological aspects of ABC
Summary statistics

Best subset selection

- **Joyce and Marjoram (2008) SAGMB**, τ-sufficiency
- **Nunes and Balding (2010) SAGMB**, entropy

Projection

- **Fearnhead and Prangle (2012) JRSS B** introduce semi-automatic ABC

Regularization techniques

- **Blum, Nunes, Prangle and Fearnhead (2013) Statistical Science** use ridge regression
- **Saulnier, Gascuel, Alizon (2017) Plos Computational Biology** use LASSO

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \ldots, M$
   a) Generate m_i from the prior $\pi(\mathscr{M} = m)$
   b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$
   c) Generate z from the model $f_{m_i}(\cdot|\theta'_{m_i})$
   d) Calculate $d_i = d(\eta(z), \eta(y))$
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$
3) Return the m_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \ldots, M$
   a) Generate m_i from the prior $\pi(\mathscr{M} = m)$
   b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$
   c) Generate z from the model $f_{m_i}(\cdot|\theta'_{m_i})$
   d) Calculate $d_i = d(\eta(z), \eta(y))$
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$
3) Return the m_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \ldots, M$
   a) Generate m_i from the prior $\pi(\mathscr{M} = m)$
   b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$
   c) Generate z from the model $f_{m_i}(\cdot|\theta'_{m_i})$
   d) Calculate $d_i = d(\eta(z), \eta(y))$
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$
3) Return the m_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \ldots, M$
   a) Generate m_i from the prior $\pi(\mathscr{M} = m)$
   b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$
   c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$
   d) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$
3) Return the m_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \ldots, M$
   a) Generate m_i from the prior $\pi(\mathscr{M} = m)$
   b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$
   c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$
   d) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$
3) Return the m_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \ldots, M$
   - a) Generate m_i from the prior $\pi(\mathscr{M} = m)$
   - b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$
   - c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$
   - d) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$
3) Return the m_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \dots, M$
   a) Generate m_i from the prior $\pi(\mathscr{M} = m)$
   b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$
   c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$
   d) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \dots, d_{(M)}$
3) Return the m_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \ldots, M$
    a) Generate m_i from the prior $\pi(\mathscr{M} = m)$
    b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$
    c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$
    d) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$
2) Order the distances $d_{(1)}, \ldots, d_{(M)}$
3) Return the m_i's that correspond to the N-smallest distances

$N = \lfloor \alpha M \rfloor$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \dots, M$

a) Generate m_i from the prior $\pi(\mathscr{M} = m)$

b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$

c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$

d) Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$

2) Order the distances $d_{(1)}, \dots, d_{(M)}$

3) Return the m_i's that correspond to the N-smallest distances

$$N = \lfloor \alpha M \rfloor$$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

1) For $i = 1, \dots, M$

- **a)** Generate m_i from the prior $\pi(\mathscr{M} = m)$
- **b)** Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$
- **c)** Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$
- **d)** Calculate $d_i = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$

2) Order the distances $d_{(1)}, \dots, d_{(M)}$

3) Return the m_i's that correspond to the N-smallest distances

$$N = \lfloor \alpha M \rfloor$$

A k-NN approximation of the posterior probabilities

Methodological aspects of ABC
ABC model choice procedure

If $\eta(\mathbf{y})$ is a sufficient statistics for the model choice problem, this can work pretty well

ABC likelihood-free methods for model choice in Gibbs random fields Grelaud, Robert, Marin, Rodolphe and Taly (2009) Bayesian Analysis

If not...

Lack of confidence in approximate Bayesian computation model choice Robert, Cornuet, Marin, Pillai (2011) PNAS

Relevant statistics for Bayesian model choice Marin, Pillai, Robert, Rousseau (2014) JRSS B

Methodological aspects of ABC
ABC model choice procedure

If $\eta(\mathbf{y})$ is a sufficient statistics for the model choice problem, this can work pretty well

ABC likelihood-free methods for model choice in Gibbs random fields Grelaud, Robert, Marin, Rodolphe and Taly (2009) Bayesian Analysis

If not...

Lack of confidence in approximate Bayesian computation model choice Robert, Cornuet, Marin, Pillai (2011) PNAS

Relevant statistics for Bayesian model choice Marin, Pillai, Robert, Rousseau (2014) JRSS B

Methodological aspects of ABC
ABC model choice procedure

If $\eta(\mathbf{y})$ is a sufficient statistics for the model choice problem, this can work pretty well

ABC likelihood-free methods for model choice in Gibbs random fields Grelaud, Robert, Marin, Rodolphe and Taly (2009) Bayesian Analysis

If not...

Lack of confidence in approximate Bayesian computation model choice Robert, Cornuet, Marin, Pillai (2011) PNAS

Relevant statistics for Bayesian model choice Marin, Pillai, Robert, Rousseau (2014) JRSS B

Methodological aspects of ABC
ABC model choice procedure

If $\eta(\mathbf{y})$ is a sufficient statistics for the model choice problem, this can work pretty well

ABC likelihood-free methods for model choice in Gibbs random fields Grelaud, Robert, Marin, Rodolphe and Taly (2009) Bayesian Analysis

If not...

Lack of confidence in approximate Bayesian computation model choice Robert, Cornuet, Marin, Pillai (2011) PNAS

Relevant statistics for Bayesian model choice Marin, Pillai, Robert, Rousseau (2014) JRSS B

Methodological aspects of ABC
ABC model choice procedure

If $\eta(\mathbf{y})$ is a sufficient statistics for the model choice problem, this can work pretty well

ABC likelihood-free methods for model choice in Gibbs random fields Grelaud, Robert, Marin, Rodolphe and Taly (2009) Bayesian Analysis

If not...

Lack of confidence in approximate Bayesian computation model choice Robert, Cornuet, Marin, Pillai (2011) PNAS

Relevant statistics for Bayesian model choice Marin, Pillai, Robert, Rousseau (2014) JRSS B

We investigate some ABC model choice techniques that use others machine learning procedures

Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics Estoup, Lombaert, Marin, Guillemaud, Pudlo, Robert, Cornuet (2012) Molecular Ecology

We investigate some ABC model choice techniques that use others machine learning procedures

Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics Estoup, Lombaert, Marin, Guillemaud, Pudlo, Robert, Cornuet (2012) Molecular Ecology

Methodological aspects of ABC
Sofwares

abc R package several ABC algorithms for performing parameter estimation and model selection

abctools R package tuning ABC analyses
https://journal.r-project.org/archive/2015-2/nunes-prangle.pdf

abcrf R package ABC via random forests

EasyABC R package several algorithms for performing efficient ABC sampling schemes, including 4 sequential sampling schemes and 3 MCMC schemes

Methodological aspects of ABC
Sofwares

DIY-ABC software performs parameter estimation and model selection for population genetics models

ABC-SysBio python package parameter inference and model selection for dynamical systems

ABCtoolbox programs various ABC algorithms including rejection sampling, MCMC without likelihood, a particle-based sampler, and ABC-GLM

PopABC software package for inference of the pattern of demographic divergence, coalescent simulation, bayesian model choice

Methodological aspects of ABC
Sofwares

Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, Estoup (2008) Bioinformatics

DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data Cornuet, Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin, Estoup (2014) Bioinformatics

DIYABC

Version 2 beta

A computer software to make inference on population evolutionary history using genetic data (microsatellites, DNA sequences and SNPs) obtained from population samples

http://www.montpellier.inra.fr/CBGP/diyabc/

Asian ladybug
European honey bee
drosophila suzukii
Pigmies populations
Four human populations, to study the out-of-Africa colonization

Methodological aspects of ABC
Sofwares

Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, Estoup (2008) Bioinformatics

DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data Cornuet, Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin, Estoup (2014) Bioinformatics

DIYABC

Version 2 beta

A computer software to make inference on population evolutionary history using genetic data (microsatellites, DNA sequences and SNPs) obtained from population samples

http://www.montpellier.inra.fr/CBGP/diyabc/

Asian ladybug
European honey bee
drosophila suzukii
Pigmies populations
Four human populations, to study the out-of-Africa colonization

Methodological aspects of ABC
Sofwares

Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation **Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, Estoup (2008) Bioinformatics**

DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data **Cornuet, Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin, Estoup (2014) Bioinformatics**

DIYABC

Version 2 beta

A computer software to make inference on population evolutionary history using genetic data (microsatellites, DNA sequences and SNPs) obtained from population samples

http://www.montpellier.inra.fr/CBGP/diyabc/

Asian ladybug
European honey bee
drosophila suzukii
Pigmies populations
Four human populations, to study the out-of-Africa colonization

Methodological aspects of ABC
Sofwares

Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, Estoup (2008) Bioinformatics

DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data Cornuet, Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin, Estoup (2014) Bioinformatics

DIYABC

Version 2 beta

A computer software to make inference on population evolutionary history using genetic data (microsatellites, DNA sequences and SNPs) obtained from population samples

http://www.montpellier.inra.fr/CBGP/diyabc/

Asian ladybug
European honey bee
drosophila suzukii
Pigmies populations
Four human populations, to study the out-of-Africa colonization

Methodological aspects of ABC
Sofwares

Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, Estoup (2008) Bioinformatics

DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data Cornuet, Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin, Estoup (2014) Bioinformatics

DIYABC

Version 2 beta

A computer software to make inference on population evolutionary history using genetic data (microsatellites, DNA sequences and SNPs) obtained from population samples

http://www.montpellier.inra.fr/CBGP/diyabc/

Asian ladybug
European honey bee
drosophila suzukii
Pigmies populations
Four human populations, to study the out-of-Africa colonization

Methodological aspects of ABC
Sofwares

Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, Estoup (2008) Bioinformatics

DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data Cornuet, Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin, Estoup (2014) Bioinformatics

DIYABC

Version 2 beta

A computer software to make inference on population evolutionary history using genetic data (microsatellites, DNA sequences and SNPs) obtained from population samples

http://www.montpellier.inra.fr/CBGP/diyabc/

Asian ladybug
European honey bee
drosophila suzukii
Pigmies populations
Four human populations, to study the out-of-Africa colonization

Methodological aspects of ABC
Sofwares

Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, Estoup (2008) Bioinformatics

DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data Cornuet, Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin, Estoup (2014) Bioinformatics

DIYABC

Version 2 beta

A computer software to make inference on population evolutionary history using genetic data (microsatellites, DNA sequences and SNPs) obtained from population samples

http://www.montpellier.inra.fr/CBGP/diyabc/

Asian ladybug
European honey bee
drosophila suzukii
Pigmies populations
Four human populations, to study the out-of-Africa colonization

Methodological aspects of ABC
Frontline news from population geneticists country

DIYABC (2014) paper has now around 500 citations

- simulate from the model can be very computationally intensive, parallelizable algorithms are necessary
- likelihoods are intractable due to the strong and complex dependence structure of the model
- sequential methods are difficult to calibrate and do not give reproducible results
- post hoc adjustments are crucial but they underestimate the amount of uncertainty
- available techniques to select the summary statistics do not give reproducible results

Methodological aspects of ABC
Frontline news from population geneticists country

DIYABC (2014) paper has now around 500 citations

- simulate from the model can be very computationally intensive, parallelizable algorithms are necessary
- likelihoods are intractable due to the strong and complex dependence structure of the model
- sequential methods are difficult to calibrate and do not give reproducible results
- post hoc adjustments are crucial but they underestimate the amount of uncertainty
- available techniques to select the summary statistics do not give reproducible results

Methodological aspects of ABC
Frontline news from population geneticists country

DIYABC (2014) paper has now around 500 citations

- simulate from the model can be very computationally intensive, parallelizable algorithms are necessary
- likelihoods are intractable due to the strong and complex dependence structure of the model
- sequential methods are difficult to calibrate and do not give reproducible results
- post hoc adjustments are crucial but they underestimate the amount of uncertainty
- available techniques to select the summary statistics do not give reproducible results

Methodological aspects of ABC
Frontline news from population geneticists country

DIYABC (2014) paper has now around 500 citations

- simulate from the model can be very computationally intensive, parallelizable algorithms are necessary
- likelihoods are intractable due to the strong and complex dependence structure of the model
- sequential methods are difficult to calibrate and do not give reproducible results
- post hoc adjustments are crucial but they underestimate the amount of uncertainty
- available techniques to select the summary statistics do not give reproducible results

Methodological aspects of ABC
Frontline news from population geneticists country

DIYABC (2014) paper has now around 500 citations

- simulate from the model can be very computationally intensive, parallelizable algorithms are necessary
- likelihoods are intractable due to the strong and complex dependence structure of the model
- sequential methods are difficult to calibrate and do not give reproducible results
- post hoc adjustments are crucial but they underestimate the amount of uncertainty
- available techniques to select the summary statistics do not give reproducible results

Methodological aspects of ABC
Frontline news from population geneticists country

DIYABC (2014) paper has now around 500 citations

- simulate from the model can be very computationally intensive, parallelizable algorithms are necessary
- likelihoods are intractable due to the strong and complex dependence structure of the model
- sequential methods are difficult to calibrate and do not give reproducible results
- post hoc adjustments are crucial but they underestimate the amount of uncertainty
- available techniques to select the summary statistics do not give reproducible results

Methodological aspects of ABC
Frontline news from population geneticists country

DIYABC (2014) paper has now around 500 citations

- simulate from the model can be very computationally intensive, parallelizable algorithms are necessary
- likelihoods are intractable due to the strong and complex dependence structure of the model
- sequential methods are difficult to calibrate and do not give reproducible results
- post hoc adjustments are crucial but they underestimate the amount of uncertainty
- available techniques to select the summary statistics do not give reproducible results

Methodological aspects of ABC
Frontline news from population geneticists country

Despite all these works, two major difficulties

- to ensure reliability of the method, the number of simulations should be large
- choice of the summaries statistics is still a problem

Methodological aspects of ABC
Frontline news from population geneticists country

Despite all these works, two major difficulties

- to ensure reliability of the method, the number of simulations should be large
- choice of the summaries statistics is still a problem

Methodological aspects of ABC
Frontline news from population geneticists country

Despite all these works, two major difficulties

- to ensure reliability of the method, the number of simulations should be large
- choice of the summaries statistics is still a problem

Methodological aspects of ABC
Use modern machine learning tools

Exploiting a large number of summary statistics is not an issue for some machine learning methods

Idea: learn on a huge reference table using random forests

Some theoretical guarantees for sparse problems

Analysis of a random forest model
Biau (2012) JMLR

Consistency of random forests
Scornet, Biau, Vert (2015) The Annals of Statistics

Methodological aspects of ABC
Use modern machine learning tools

Exploiting a large number of summary statistics is not an issue for some machine learning methods

Idea: learn on a huge reference table using random forests

Some theoretical guarantees for sparse problems

Analysis of a random forest model
Biau (2012) JMLR

Consistency of random forests
Scornet, Biau, Vert (2015) The Annals of Statistics

Methodological aspects of ABC
Use modern machine learning tools

Exploiting a large number of summary statistics is not an issue for some machine learning methods

Idea: learn on a huge reference table using random forests

Some theoretical guarantees for sparse problems

Analysis of a random forest model
Biau (2012) JMLR

Consistency of random forests
Scornet, Biau, Vert (2015) The Annals of Statistics

Methodological aspects of ABC
Use modern machine learning tools

Exploiting a large number of summary statistics is not an issue for some machine learning methods

Idea: learn on a huge reference table using random forests

Some theoretical guarantees for sparse problems

Analysis of a random forest model
Biau (2012) JMLR

Consistency of random forests
Scornet, Biau, Vert (2015) The Annals of Statistics

Methodological aspects of ABC
Use modern machine learning tools

Exploiting a large number of summary statistics is not an issue for some machine learning methods

Idea: learn on a huge reference table using random forests

Some theoretical guarantees for sparse problems

Analysis of a random forest model
Biau (2012) JMLR

Consistency of random forests
Scornet, Biau, Vert (2015) The Annals of Statistics

Methodological aspects of ABC
Use modern machine learning tools

Exploiting a large number of summary statistics is not an issue for some machine learning methods

Idea: learn on a huge reference table using random forests

Some theoretical guarantees for sparse problems

Analysis of a random forest model
Biau (2012) JMLR

Consistency of random forests
Scornet, Biau, Vert (2015) The Annals of Statistics

Methodological aspects of ABC
Use modern machine learning tools

This work stands at the interface between Bayesian inference and machine learning techniques

As an alternative, Papamakarios and Murray (2016) propose to approximate the whole posterior distribution by using Mixture Density Networks (MDN, Bishop, 1994)

Fast e-free Inference of Simulation Models with Bayesian Conditional Density Estimation
Papamakarios and Murray (2016) NIPS

Methodological aspects of ABC
Use modern machine learning tools

This work stands at the interface between Bayesian inference and machine learning techniques

As an alternative, Papamakarios and Murray (2016) propose to approximate the whole posterior distribution by using Mixture Density Networks (MDN, Bishop, 1994)

Fast e-free Inference of Simulation Models with Bayesian Conditional Density Estimation

Papamakarios and Murray (2016) NIPS

Methodological aspects of ABC
Use modern machine learning tools

This work stands at the interface between Bayesian inference and machine learning techniques

As an alternative, Papamakarios and Murray (2016) propose to approximate the whole posterior distribution by using Mixture Density Networks (MDN, Bishop, 1994)

Fast e-free Inference of Simulation Models with Bayesian Conditional Density Estimation
Papamakarios and Murray (2016) NIPS

Methodological aspects of ABC
Use modern machine learning tools

The MDN strategy consists in using Gaussian mixture models with parameters calibrated thanks to neural networks

Idea: iteratively learn an efficient proposal prior (approximating the posterior distribution), then to use this proposal to train the posterior, both steps making use of MDN

The number of mixture components and the number of hidden layers of the networks require calibration

Methodological aspects of ABC
Use modern machine learning tools

The MDN strategy consists in using Gaussian mixture models with parameters calibrated thanks to neural networks

Idea: iteratively learn an efficient proposal prior (approximating the posterior distribution), then to use this proposal to train the posterior, both steps making use of MDN

The number of mixture components and the number of hidden layers of the networks require calibration

Methodological aspects of ABC
Use modern machine learning tools

The MDN strategy consists in using Gaussian mixture models with parameters calibrated thanks to neural networks

Idea: iteratively learn an efficient proposal prior (approximating the posterior distribution), then to use this proposal to train the posterior, both steps making use of MDN

The number of mixture components and the number of hidden layers of the networks require calibration

Methodological aspects of ABC
Use modern machine learning tools

Deep Learning for Population Genetic Inference
Sheehan and Song (2016) PLOS Computational Biology

Deep learning makes use of multilayer neural networks to learn a feature-based function from the input (hundreds of correlated summary statistics) to the output (population genetic parameters of interest).

Unsupervised pretraining using autoencoders very interesting, but requires a lot of calibration

Methodological aspects of ABC
Use modern machine learning tools

Deep Learning for Population Genetic Inference
Sheehan and Song (2016) PLOS Computational Biology

Deep learning makes use of multilayer neural networks to learn a feature-based function from the input (hundreds of correlated summary statistics) to the output (population genetic parameters of interest).

Unsupervised pretraining using autoencoders very interesting, but requires a lot of calibration

Methodological aspects of ABC
Use modern machine learning tools

Deep Learning for Population Genetic Inference
Sheehan and Song (2016) PLOS Computational Biology

Deep learning makes use of multilayer neural networks to learn a feature-based function from the input (hundreds of correlated summary statistics) to the output (population genetic parameters of interest).

Unsupervised pretraining using autoencoders very interesting, but requires a lot of calibration

ABC random forests
Model choice

Reliable ABC model choice via random forests **Pudlo, Marin, Estoup, Cornuet, Gauthier and Robert (2016) Bioinformatics**

Input ABC reference table involving model index and summary statistics, table used as learning set

possibly large collection of summary statistics: from scientific theory input to machine-learning alternatives

For $i = 1, \dots, M$

Output a random forest classifier to infer model indexes $\widehat{\mathfrak{m}(\eta(\mathbf{y}))}$

ABC random forests
Model choice

Reliable ABC model choice via random forests Pudlo, Marin, Estoup, Cornuet, Gauthier and Robert (2016) Bioinformatics

Input ABC reference table involving model index and summary statistics, table used as learning set

possibly large collection of summary statistics: from scientific theory input to machine-learning alternatives

For $i = 1, \ldots, M$

Output a random forest classifier to infer model indexes $\widehat{m(\eta(y))}$

ABC random forests
Model choice

Reliable ABC model choice via random forests **Pudlo, Marin, Estoup, Cornuet, Gauthier and Robert (2016) Bioinformatics**

Input ABC reference table involving model index and summary statistics, table used as learning set

possibly large collection of summary statistics: from scientific theory input to machine-learning alternatives

```
For $i = 1, \ldots, M$
    Generate $m_i$ from the prior $\pi(\mathcal{M} = m)$
    Generate $\theta'_{m_i}$ from the prior $\pi_{m_i}(\cdot)$
    Generate $z$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$
    Calculate $x_i = \eta(z)$
```

Output a random forest classifier to infer model indexes $m\widehat{(\eta(y))}$

ABC random forests
Model choice

Reliable ABC model choice via random forests **Pudlo, Marin, Estoup, Cornuet, Gauthier and Robert (2016) Bioinformatics**

Input ABC reference table involving model index and summary statistics, table used as learning set

possibly large collection of summary statistics: from scientific theory input to machine-learning alternatives

For $i = 1, \dots, M$

a) Generate m_i from the prior $\pi(\mathscr{M} = m)$

b) Generate θ'_{m_i} from the prior $\pi_{m_i}(\cdot)$

c) Generate z from the model $f_{m_i}(\cdot|\theta'_{m_i})$

d) Calculate $x_i = \eta(z_i)$

Output a random forest classifier to infer model indexes $\widehat{m(\eta(\mathbf{y}))}$

ABC random forests
Model choice

Reliable ABC model choice via random forests **Pudlo, Marin, Estoup, Cornuet, Gauthier and Robert (2016) Bioinformatics**

Input ABC reference table involving model index and summary statistics, table used as learning set

possibly large collection of summary statistics: from scientific theory input to machine-learning alternatives

```
For $i = 1, \dots, M$
  a) Generate $m_i$ from the prior $\pi(\mathscr{M} = m)$
  b) Generate $\theta'_{m_i}$ from the prior $\pi_{m_i}(\cdot)$
  c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$
  d) Calculate $\mathbf{x}_i = \eta(\mathbf{z}_i)$
```

Output a random forest classifier to infer model indexes $\widehat{m(\eta(\mathbf{y}))}$

ABC random forests
Model choice

Reliable ABC model choice via random forests **Pudlo, Marin, Estoup, Cornuet, Gauthier and Robert (2016) Bioinformatics**

Input ABC reference table involving model index and summary statistics, table used as learning set

possibly large collection of summary statistics: from scientific theory input to machine-learning alternatives

```
For $i = 1, \ldots, M$
  a) Generate $m_i$ from the prior $\pi(\mathscr{M} = m)$
  b) Generate $\theta'_{m_i}$ from the prior $\pi_{m_i}(\cdot)$
  c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\theta'_{m_i})$
  d) Calculate $\mathbf{x}_i = \eta(\mathbf{z}_i)$
```

Output a random forest classifier to infer model indexes $\widehat{m(\eta(\mathbf{y}))}$

ABC random forests
Model choice

Random forest predicts a MAP model index, from the observed dataset

the predictor provided by the forest is good enough to select the most likely model

but not to derive directly the associated posterior probabilities

frequency of trees associated with majority model is no proper substitute to the true posterior probability

ABC random forests
Model choice

Random forest predicts a MAP model index, from the observed dataset

the predictor provided by the forest is good enough to select the most likely model

but not to derive directly the associated posterior probabilities

frequency of trees associated with majority model is no proper substitute to the true posterior probability

ABC random forests
Model choice

Random forest predicts a MAP model index, from the observed dataset

the predictor provided by the forest is good enough to select the most likely model

but not to derive directly the associated posterior probabilities

frequency of trees associated with majority model is no proper substitute to the true posterior probability

ABC random forests
Model choice

Random forest predicts a MAP model index, from the observed dataset

the predictor provided by the forest is good enough to select the most likely model

but not to derive directly the associated posterior probabilities

frequency of trees associated with majority model is no proper substitute to the true posterior probability

ABC random forests
Model choice

Estimate of the posterior probability of the selected model

$$\mathbb{P}[\mathscr{M} = \widehat{m(\eta(\mathbf{y}))}|\eta(\mathbf{y})]$$

random comes from $\mathscr{M}$ (bayesian)!

$$\mathbb{P}[\mathscr{M} = \widehat{m(\eta(\mathbf{y}))}|\eta(\mathbf{y})] = 1 - \mathbb{E}\left[\mathbb{I}(\mathscr{M} \neq \widehat{m(\eta(\mathbf{y}))})|\eta(\mathbf{y})\right]$$

ABC random forests
Model choice

Estimate of the posterior probability of the selected model

$$\mathbb{P}[\mathscr{M} = \widehat{m(\eta(\mathbf{y}))} | \eta(\mathbf{y})]$$

random comes from $\mathscr{M}$ (bayesian)!

$$\mathbb{P}[\mathscr{M} = \widehat{m(\eta(\mathbf{y}))} | \eta(\mathbf{y})] = 1 - \mathbb{E}\left[\mathbb{I}(\mathscr{M} \neq \widehat{m(\eta(\mathbf{y}))}) | \eta(\mathbf{y})\right]$$

A second random forest in regression

1) compute the value of $\mathbb{I}(\mathscr{M} \neq \widehat{\mathfrak{m}(\eta(\mathbf{z}))})$ for the trained random forest $\hat{\mathfrak{m}}$ and for all terms in the ABC reference table using the out-of-bag classifiers
2) train a RF regression and get $\widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \widehat{\mathfrak{m}(\eta(\mathbf{z}))})|\eta(\mathbf{z})\right]$
3) return $\widehat{\mathbb{P}}[\mathscr{M} = \widehat{\mathfrak{m}(\eta(\mathbf{y}))}|\eta(\mathbf{y})] = 1 - \widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \widehat{\mathfrak{m}(\eta(\mathbf{z}))})|\eta(\mathbf{z})\right]$

on same reference table out-of-bag magic trick avoid over-fitting!

ABC random forests
Model choice

A second random forest in regression

1) compute the value of $\mathbb{I}(\mathscr{M} \neq \mathrm{m}\widehat{(\eta(\mathbf{z}))})$ for the trained random forest $\hat{\mathrm{m}}$ and for all terms in the ABC reference table using the out-of-bag classifiers

2) train a RF regression and get $\widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \mathrm{m}\widehat{(\eta(\mathbf{z}))})|\eta(\mathbf{z})\right]$

3) return
$\widehat{\mathbb{P}}[\mathscr{M} = \mathrm{m}\widehat{(\eta(\mathbf{y}))}|\eta(\mathbf{y})] = 1 - \widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \mathrm{m}\widehat{(\eta(\mathbf{z}))})|\eta(\mathbf{z})\right]$

on same reference table out-of-bag magic trick avoid overfitting!

ABC random forests
Model choice

A second random forest in regression

1) compute the value of $\mathbb{I}(\mathscr{M} \neq \widehat{\mathrm{m}(\eta(\mathbf{z})}))$ for the trained random forest $\hat{\mathrm{m}}$ and for all terms in the ABC reference table using the out-of-bag classifiers

2) train a RF regression and get $\widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \mathrm{m}(\widehat{\eta(\mathbf{z})}))|\eta(\mathbf{z})\right]$

3) return
$\widehat{\mathbb{P}}[\mathscr{M} = \mathrm{m}(\widehat{\eta(\mathbf{y})})|\eta(\mathbf{y})] = 1 - \widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \mathrm{m}(\widehat{\eta(\mathbf{z})}))|\eta(\mathbf{z})\right]$

on same reference table out-of-bag magic trick avoid overfitting!

ABC random forests
Model choice

A second random forest in regression

1) compute the value of $\mathbb{I}(\mathscr{M} \neq \widehat{\mathrm{m}(\eta(\mathbf{z}))})$ for the trained random forest $\hat{\mathrm{m}}$ and for all terms in the ABC reference table using the out-of-bag classifiers
2) train a RF regression and get $\widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \widehat{\mathrm{m}(\eta(\mathbf{z}))})|\eta(\mathbf{z})\right]$
3) return
$$\widehat{\mathbb{P}}[\mathscr{M} = \widehat{\mathrm{m}(\eta(\mathbf{y}))}|\eta(\mathbf{y})] = 1 - \widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \widehat{\mathrm{m}(\eta(\mathbf{z}))})|\eta(\mathbf{z})\right]$$

on same reference table out-of-bag magic trick avoid overfitting!

ABC random forests
Model choice

A second random forest in regression

1) compute the value of $\mathbb{I}(\mathscr{M} \neq \widehat{\mathrm{m}(\eta(\mathbf{z})})$ for the trained random forest $\hat{\mathrm{m}}$ and for all terms in the ABC reference table using the out-of-bag classifiers
2) train a RF regression and get $\widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \widehat{\mathrm{m}(\eta(\mathbf{z})}))|\eta(\mathbf{z})]\right]$
3) return
$$\widehat{\mathbb{P}}[\mathscr{M} = \widehat{\mathrm{m}(\eta(\mathbf{y})})|\eta(\mathbf{y})] = 1 - \widehat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \widehat{\mathrm{m}(\eta(\mathbf{z})}))|\eta(\mathbf{z})]\right]$$

on same reference table out-of-bag magic trick avoid overfitting!

ABC random forests
Parameter inference

ABC random forests for Bayesian parameter inference **Raynal, Marin, Pudlo, Ribatet, Robert and Estoup (2017) Preprint reviewed and recommended by Peer Community In Evolutionary Biology**

Input ABC reference table involving parameters values and summary statistics, table used as learning set

```
    For $i = 1, \ldots, M$
```

Output some regression RF predictors to infer posterior expectations, quantiles, variances and covariances

ABC random forests
Parameter inference

ABC random forests for Bayesian parameter inference **Raynal, Marin, Pudlo, Ribatet, Robert and Estoup (2017) Preprint reviewed and recommended by Peer Community In Evolutionary Biology**

Input ABC reference table involving parameters values and summary statistics, table used as learning set

```
For $i = 1, \ldots, M$
    Generate $\theta_i$ from the prior $\pi(\cdot)$
    Generate $z_i$ from the model $f(\cdot|\theta_i)$
    Calculate $x_i = \eta(z_i)$
```

Output some regression RF predictors to infer posterior expectations, quantiles, variances and covariances

ABC random forests
Parameter inference

ABC random forests for Bayesian parameter inference Raynal, Marin, Pudlo, Ribatet, Robert and Estoup (2017) Preprint reviewed and recommended by Peer Community In Evolutionary Biology

Input ABC reference table involving parameters values and summary statistics, table used as learning set

```
For $i = 1, \ldots, M$
    a) Generate $\theta_i$ from the prior $\pi(\cdot)$
    b) Generate $z_i$ from the model $f(\cdot|\theta_i)$
    c) Calculate $x_i = \eta(z_i)$
```

Output some regression RF predictors to infer posterior expectations, quantiles, variances and covariances

ABC random forests
Parameter inference

ABC random forests for Bayesian parameter inference Raynal, Marin, Pudlo, Ribatet, Robert and Estoup (2017) Preprint reviewed and recommended by Peer Community In Evolutionary Biology

Input ABC reference table involving parameters values and summary statistics, table used as learning set

```
For $i = 1, \dots, M$
  a) Generate $\theta_i$ from the prior $\pi(\cdot)$
  b) Generate $\mathbf{z}_i$ from the model $f(\cdot|\theta_i)$
  c) Calculate $\mathbf{x}_i = \eta(\mathbf{z}_i)$
```

Output some regression RF predictors to infer posterior expectations, quantiles, variances and covariances

ABC random forests
Parameter inference

ABC random forests for Bayesian parameter inference Raynal, Marin, Pudlo, Ribatet, Robert and Estoup (2017) Preprint reviewed and recommended by Peer Community In Evolutionary Biology

Input ABC reference table involving parameters values and summary statistics, table used as learning set

For $i = 1, \dots, M$

a) Generate θ_i from the prior $\pi(\cdot)$

b) Generate $\mathbf{z}_i$ from the model $f(\cdot|\theta_i)$

c) Calculate $\mathbf{x}_i = \eta(\mathbf{z}_i)$

Output some regression RF predictors to infer posterior expectations, quantiles, variances and covariances

ABC random forests
Parameter inference

Expectations Construct d regression RF, one per dimension

Quantiles very nice trick to estimate the cdf, no new forest
Quantile Regression Forests Meinshausen (2006) JMLR

Variances use of a out-of-bag trick, no new forest

Covariances new forests for which the responses variables are the products of out-of-bag errors

ABC random forests
Parameter inference

Expectations Construct d regression RF, one per dimension

Quantiles very nice trick to estimate the cdf, no new forest
Quantile Regression Forests Meinshausen (2006) JMLR

Variances use of a out-of-bag trick, no new forest

Covariances new forests for which the responses variables are the products of out-of-bag errors

ABC random forests
Parameter inference

Expectations Construct d regression RF, one per dimension

Quantiles very nice trick to estimate the cdf, no new forest
Quantile Regression Forests **Meinshausen (2006) JMLR**

Variances use of a out-of-bag trick, no new forest

Covariances new forests for which the responses variables are the products of out-of-bag errors

ABC random forests
Parameter inference

Expectations Construct d regression RF, one per dimension

Quantiles very nice trick to estimate the cdf, no new forest
Quantile Regression Forests **Meinshausen (2006) JMLR**

Variances use of a out-of-bag trick, no new forest

Covariances new forests for which the responses variables are the products of out-of-bag errors

ABC random forests
Parameter inference

We constructed forests able to estimate everywhere in the space of summary statistics but we are interested only in one point, the observed dataset

construct local random forest, thesis of Louis Raynal

ABC random forests
Parameter inference

We constructed forests able to estimate everywhere in the space of summary statistics but we are interested only in one point, the observed dataset

construct local random forest, thesis of Louis Raynal

Human populations example

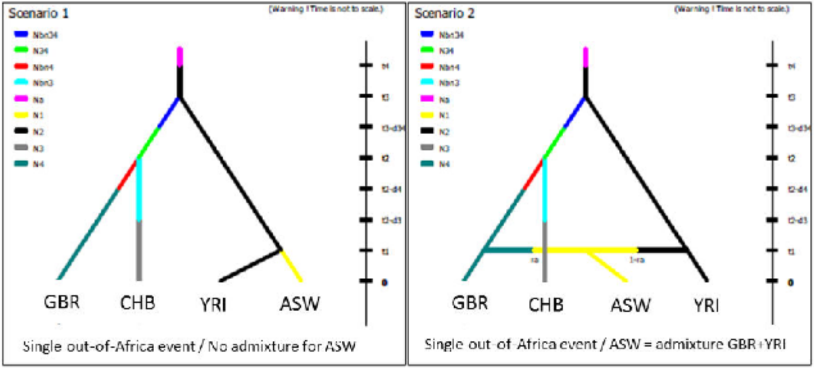
50,000 SNP markers genotyped in four Human populations: Yoruba (Africa), Han (East Asia), British (Europe) and American individuals of African Ancestry; 30 individuals per population.

We compared six scenarios of evolution which differ from each other by one ancient and one recent historical events:

A) a single out-of-Africa colonization event giving an ancestral out-of-Africa versus two independent out-of-Africa colonization events;

B) the possibility of a recent genetic admixture of Americans of African origin with their African ancestors and individuals of European or East Asia origins.

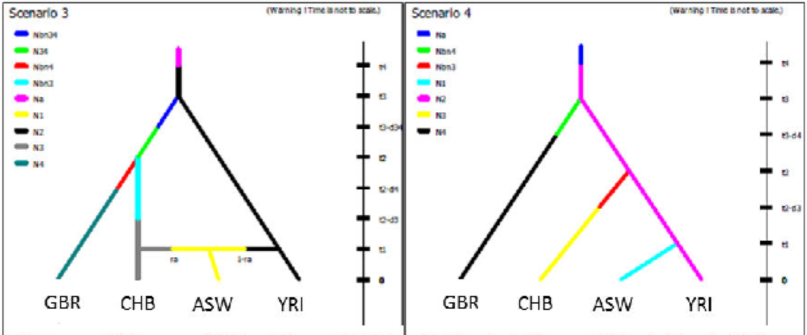
Human populations example



Single out-of-Africa event / No admixture for ASW

Single out-of-Africa event / ASW = admixture GBR+YRI

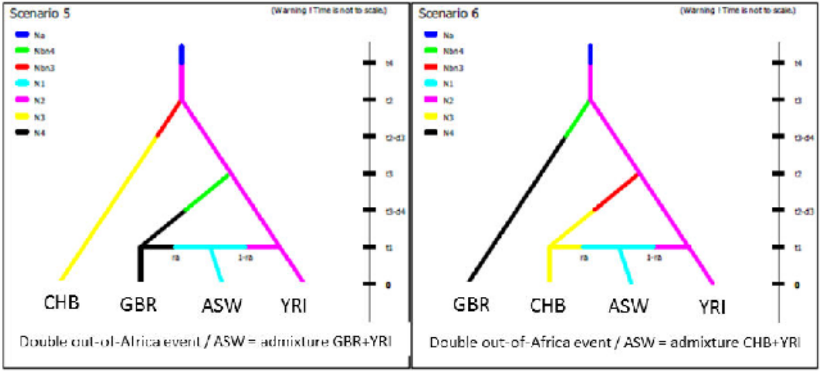
Human populations example



Single out of Africa events / ASW = admixture CHB+YRI

Double out-of-Africa event / No admixture for ASW

Human populations example



Double out-of-Africa event / ASW = admixture GBR+YRI

Double out-of-Africa event / ASW = admixture CHB+YRI

Human populations example

$\mathrm{d} = 112$ summary statistics provided by DIYABC for SNP markers complemented by the five Linear Discriminant Analysis axes

$M = 50,000$

ABC-RF algorithm selects scenario 2

With second regression forest, we got an estimate of the posterior probability of scenario 2 equal to 0.998

Human populations example

$\mathrm{d} = 112$ summary statistics provided by DIYABC for SNP markers complemented by the five Linear Discriminant Analysis axes

$M = 50,000$

ABC-RF algorithm selects scenario 2

With second regression forest, we got an estimate of the posterior probability of scenario 2 equal to 0.998

Human populations example

$\mathrm{d} = 112$ summary statistics provided by DIYABC for SNP markers complemented by the five Linear Discriminant Analysis axes

$\mathrm{M} = 50,000$

ABC-RF algorithm selects scenario 2

With second regression forest, we got an estimate of the posterior probability of scenario 2 equal to 0.998

Human populations example

$\mathrm{d} = 112$ summary statistics provided by DIYABC for SNP markers complemented by the five Linear Discriminant Analysis axes

$\mathrm{M} = 50,000$

ABC-RF algorithm selects scenario 2

With second regression forest, we got an estimate of the posterior probability of scenario 2 equal to 0.998

Human populations example

Considering previous population genetics studies in the field, it is not surprising we got

- a single out-of-Africa colonization event giving an ancestral out-of- Africa population
- a secondarily split into one European and one East Asian population lineage
- a recent genetic admixture of Americans of African origin with their African ancestors and European

Human populations example

Considering previous population genetics studies in the field, it is not surprising we got

- a single out-of-Africa colonization event giving an ancestral out-of- Africa population
- a secondarily split into one European and one East Asian population lineage
- a recent genetic admixture of Americans of African origin with their African ancestors and European

Human populations example

Considering previous population genetics studies in the field, it is not surprising we got

- a single out-of-Africa colonization event giving an ancestral out-of- Africa population
- a secondarily split into one European and one East Asian population lineage
- a recent genetic admixture of Americans of African origin with their African ancestors and European

Human populations example

Considering previous population genetics studies in the field, it is not surprising we got

- a single out-of-Africa colonization event giving an ancestral out-of- Africa population
- a secondarily split into one European and one East Asian population lineage
- a recent genetic admixture of Americans of African origin with their African ancestors and European