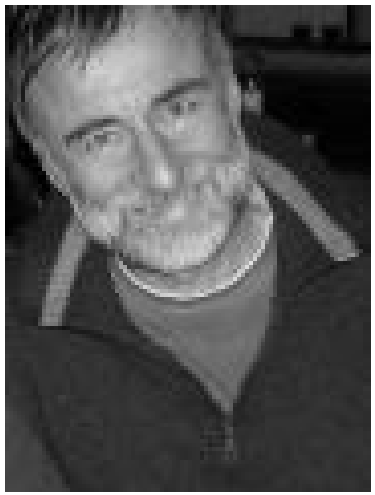
Approximate Bayesian model choice as a Machine Learning problem

Pierre Pudlo

Aix-Marseille University
Institut de Mathématiques de Marseille (I2M)

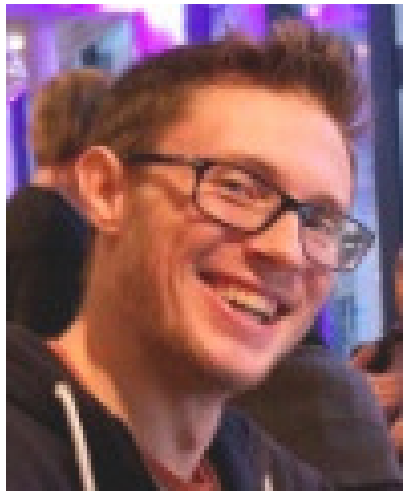
November 27th, 2018

Joint work with



Jean-Marie Cornuet, Arnaud Estoup and Mathieu Gautier






J.-M. Marin, C.P. Robert, J. Stoehr and G. Aufort

Table of Contents

Table of Contents

Bayesian model choice

The evidence of model $\mathscr{M}$

- is $p(\mathbf{x}|\mathscr{M}) =$

$$\int \pi(\theta|\mathscr{M}) f(\mathscr{P}|\theta,\mathscr{M}) f(\mathbf{x}|\mathscr{P},\theta,\mathscr{M}) \,\mathrm{d}\mathscr{P}\mathrm{d}\theta$$

where:

- $\mathscr{P}$ is the past history
- $\mathbf{x}$ is data collected at present
- $\mathscr{M}$ is the model index
- θ the parameters

Prior/posterior on the collection of models

- prior probability of model $\mathscr{M}$: $p(\mathscr{M})$
- posterior probability of model $\mathscr{M}$: $p(\mathscr{M}|x) \propto p(\mathscr{M})p(\mathbf{x}|\mathscr{M})$

The MAP model

- selects the model with maximum a posteriori probability

How to compute the posterior probability?

or the MAP model?

Approximate Bayesian computation (ABC)

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

$$f(\mathbf{x}|\theta)$$

- is (really!) not available in closed form
- cannot (easily!) be either completed or demarginalised
- cannot be (at all!) estimated by an unbiased estimator

Issue. Prohibits direct implementation of a generic MCMC algorithm like Metropolis-Hastings, Gibbs, (or EM algorithms)

In population genetics: a latent process

- of high dimension,
- including combinatorial structures

$\Longrightarrow$ intractable likelihoods

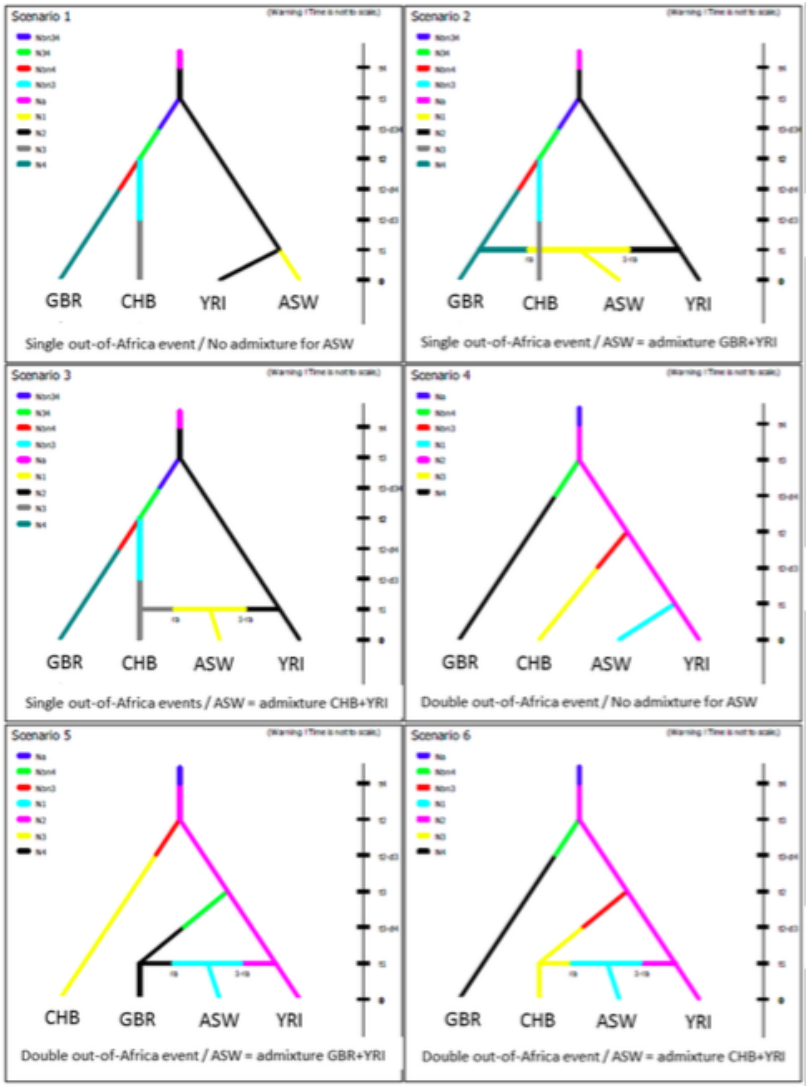
ABC is a computational technique that only requires being able to sample from the likelihood $f(\mathbf{x}|\theta)$.
Griffiths et al. (1997); Tavaré et al. (1999)

Getting approximative

- Summarising/replacing the data with (possibly insufficient) statistics
- Replacing the likelihood with a nonparametric approximation

A Model choice issue in population genetics

History of the Human population



Data

- with samples from
  - Nigeria (YRI)
  - China (CHB)
  - England (GBR)
  - African-Americans (ASW)
- genotyped at $L = 50,000$ loci on the autosomal chromosomes

Questions?

- A single out-of-Africa colonization event? or two?
- Can ASW be explained by admixture between GBR & YRI?

Dimension of θ depends on the model

ABC in Astrophysics

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

$$f(\mathbf{x}|\theta)$$

- is (really!) not available in closed form
- cannot (easily!) be either completed or demarginalised
- cannot be (at all!) estimated by an unbiased estimator

Issue. Prohibits direct implementation of a generic MCMC algorithm like Metropolis-Hastings, Gibbs, (or EM algorithms)

In Astrophysics: we have lots of datasets to analyse

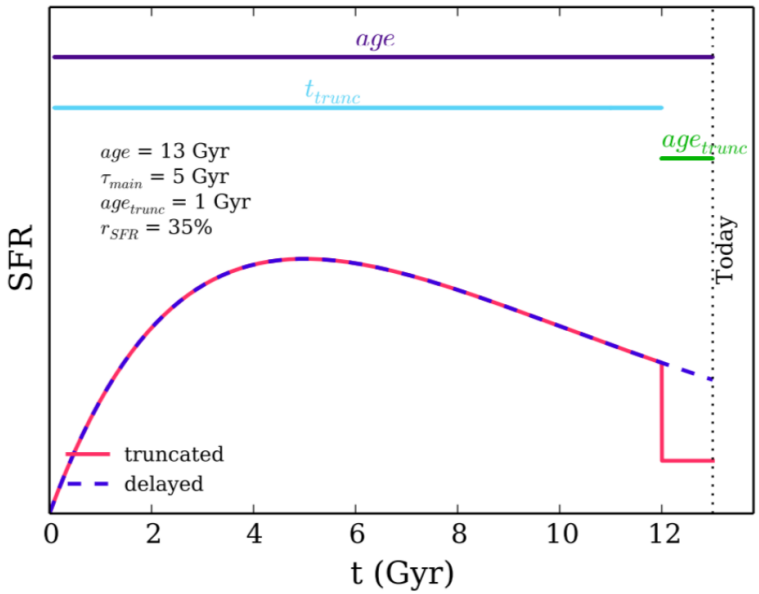
- with the same (not that intractable) likelihood
- and the same prior

$\Longrightarrow$ ABC speeds up the computation

Almost not approximative

- We do not summarise the data: each galaxy is represented by a vector of dimension ≈ 20.
- Replacing the likelihood with a nonparametric approximation on simulations

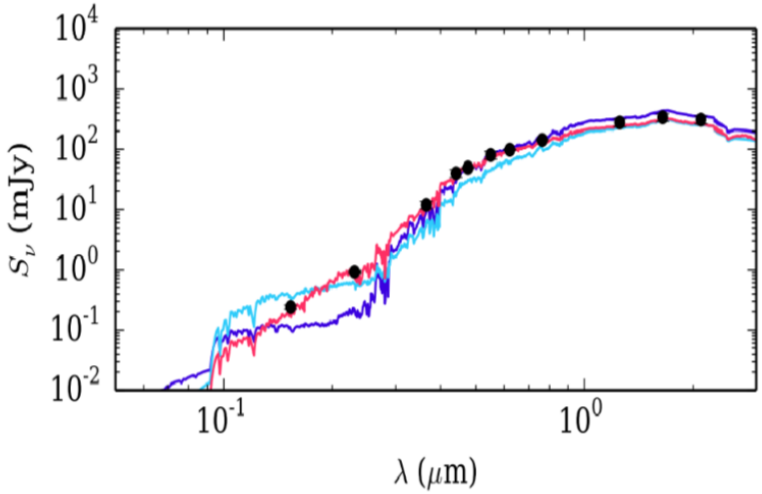
A Model choice issue in Astrophysics



Star formation history (SFH) of a galaxy

Should we had a break to account for recent variation in the SFH?

Data: Spectral Energy Distribution (SED) sampled at a few points ($\tilde{2}0$ datapoints per galaxy)



3 simulated Spectral Energy Distributions (SED)

Model: a complex model that takes SFH & many other (unknown!) parameters as entry and return a simulated SED.

Table of Contents

ABC model choice

Simulation algorithm

For i in $1 : N$

- Draw $\mathscr{M}_i$ from prior probability
- Draw θ_i from prior of the model
- Draw a dataset $\mathbf{x}_i$ from $\mathscr{M}_i, \theta_i$

EndFor

Summarize datasets

- with a non linear S : data space $\to \mathbb{R}^d$
- to compare $S(\mathbf{x}_{obs})$ & the $S(\mathbf{x}_i) = (S_1(\mathbf{x}_i), \ldots, S_d(\mathbf{x}_i))$'s

Questions?

- What can be said about $\mathscr{M}|\mathbf{x}_{\text{obs}}$ with the help of the simulations $(\mathscr{M}_i, \mathbf{x_i})$ drawn from the joint distribution?

Rejection ABC

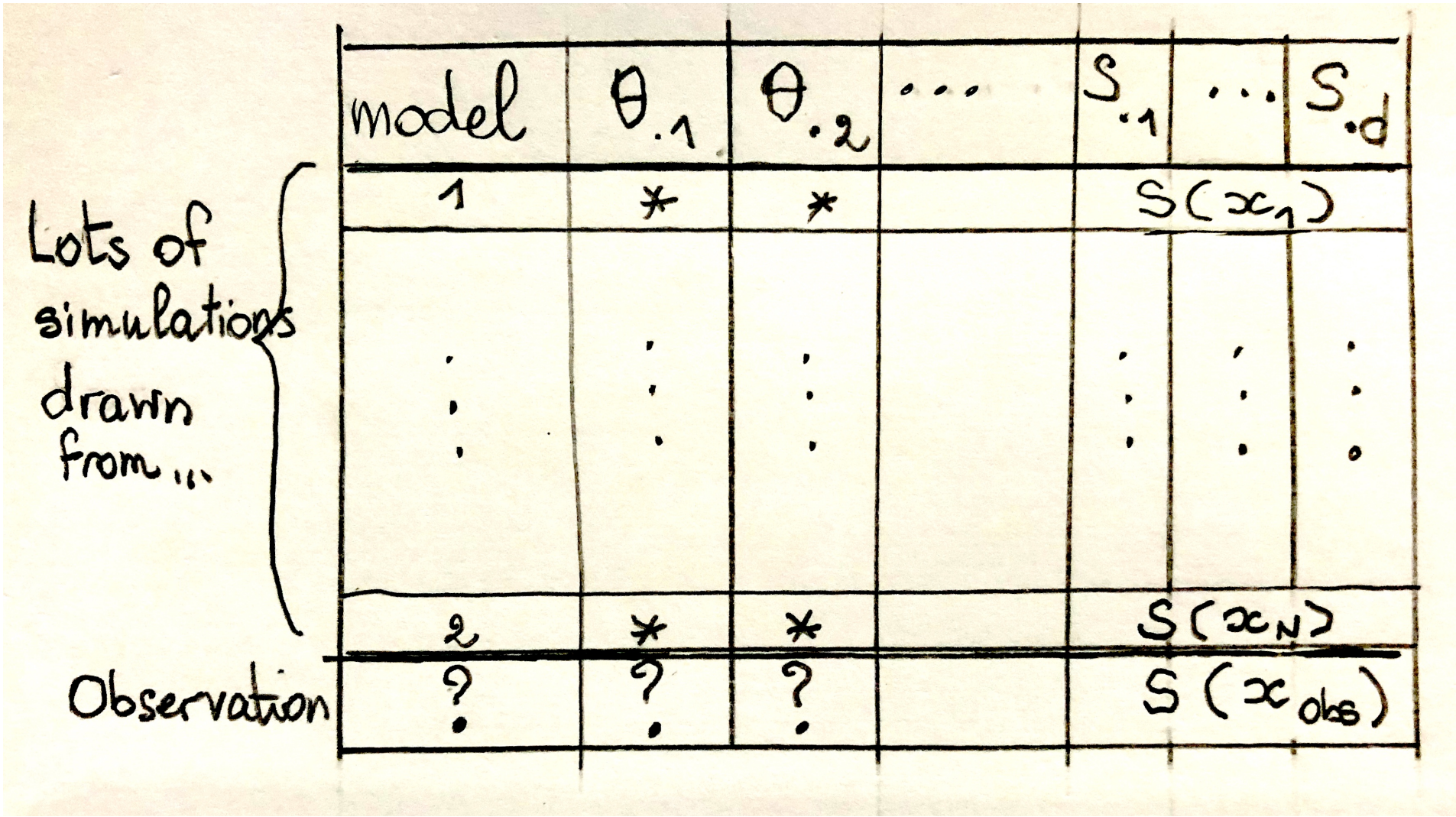
- Choose a threeshold ε
- Compute the frequency of each model among the simulations that satisfy $\|S(\mathbf{x}_i) - S(\mathbf{x}_{\text{obs}})\| \leq \varepsilon$

How to tune ε?

- Set ε so that the number of accepted simulations is K_{accepted}

$\Longrightarrow$ Looks like K-nearest neighbor method

The reference table of simulations



The Machine Learning perspective

The reference table

- A large set of N simulations $(\mathscr{M}_i, \mathbf{x}_i)$ drawn from the Bayesian model:

$$p(\mathscr{M})\pi(\theta|\mathscr{M})f(\mathscr{P}|\theta,\mathscr{M})f(\mathbf{x}|\theta,\mathscr{M})$$

Key point

$\Longrightarrow$ Train a machine learning algorithm with the N simulations:

- Response: the model index $\mathscr{M}$,
- Covariates: the summary statistics $S(\mathbf{x})$.

The goal

- Find the MAP model

$\Longleftrightarrow$

Predict the unknown $\mathscr{M}$

- Compute the posterior pr. $p(\mathscr{M} = 1|\mathbf{x}_{\text{obs}})$

$\Longleftrightarrow$

Predict the average response $p(\mathscr{M} = 1|\mathbf{x}_{\text{obs}}) = \mathbb{E}\left(\mathbf{1}\{\mathscr{M} = 1\}\middle|\mathbf{x}_{\text{obs}}\right)$

Biau, Cérou, Guyader (2015): New insights into Approx. Bayesian Comput.
Pudlo, Marin et al. (2016): Reliable ABC model choice via random forests

The ML perspective (continued)

- If the response is $\mathscr{M}$ which is discrete, we face a classification problem on the set of simulations
- If the response is the indicator vector $(0, \ldots, 1, 0, \ldots, 0)$, we face a regression problem

Basic ABC model choice

- Select the k closest $S(\mathbf{x}_i)$'s to the observed data $S(\mathbf{x})$
- Predict $\mathscr{M}$ as the most frequent model among these selected simulations
  or
- Return the frequency of each model among these simulations (=averaging the indicator vectors)

Key point

- can be interpreted as a k-nearest neighbor learning method on the set of simulations
- k is the number of selected simulations at 1st stage

Other ABC algorithms from the literature

- can be interpreted as a well-known learning method on the simulations
- Eg. Beaumont's postprocessing = local linear methods

$\Longrightarrow$ all local or nn methods suffer from the curse of dimensionality: dim d of $S(\mathbf{x})$ should be small

Our use of random forest

First random forest

- We first renounce approximating the posterior probabilities
- We begin by training a random forest on the reference table of simulations
  - to predict the model index (the response)
  - based on the summary statistics (the covariates)
- This gives us an approximation of the MAP model

$$\hat{\mathscr{M}}(\mathbf{x})$$

The prior misclassification error rate

- The amount of errors made by the random forest on simulations drawn the prior distribution

$$=$$

the prior error rate

- It represents how difficult the two models (likelihoods & priors) are separated from each other
- It can be computed (easily) with cross-validation (or out-of-bag techniques on RF)

But does the observed data fall into a part of the data space where it is difficult to assess a model? $\rightarrow$ Conditional error rate knowing $\mathbf{x}$

The second random forest

The conditional misclassification error rate knowing x

After training the first random forest,
For i in $1:N$

- compute the out-of-bag prediction $\widehat{\mathscr{M}}(\mathbf{x}_i)$ for each simulation
- set $Y_i = \mathbf{1}\{\widehat{\mathscr{M}}(\mathbf{x}_i) \neq \mathscr{M}_i\}$

EndFor

Proposition

The conditional error rate
$= 1-$ posterior pr of the MAP

$$\mathbb{E}(\mathbf{1}\{\widehat{\mathscr{M}}(\mathbf{x}) \neq \mathscr{M}\}|S(\mathbf{x})) = 1-\mathbb{P}(\widehat{\mathscr{M}}(\mathbf{x})|S(\mathbf{x}))$$

Train a second random forest

- to predict Y_i knowing $\mathbf{x}_i$
- with L^2-loss

Reliable because

- a univariate response Y
- based the best prediction of the MAP (the 1st random forest)
- the out-of-bag trick avoids underestimating the error (without resorting cross-validation)

Other Machine Learning Techniques

- Papamakarios G., Murray I (NIPS, 2016):
  Fast ε-free inference of simulation models with Bayesian conditional density estimation
- Bai Jiang, T-Y Wu, C Zheng, W H. Wong:
  ABC via Deep Neural Network (2017, arxiv)

- **In Astrophysics**, we are currently trying to use
  - Gradient Boosting Machine (XGBoost)
  - Deep Neural Network
- Instead of the RF two stages' algorithm, we train only one machine to compute directly the posterior probability of the most complex model
- With Deep NN: the main difficulty is to find a good network architecture
  $\Longrightarrow$ Grégoire Aufort
- Results in term of prior error rate where comparable ($\approx 10\%$)

Table of Contents

On the Human history

Method	prior error(%)
LDA	9.91
k-nn on $S(\mathbf{x})$	23.18
k-nn on LDA axes	6.29
RF	8.84
RF on $S(\mathbf{x})$ & LDA	5.01

on a set of $N = 10,000$ *simulations*
& $dim(S(\mathbf{x})) = 112$

- Random forest (RF) is the best classifier if we complements the original summary statistics with projections of $S(\mathbf{x}_i)$ on the axes of a linear discriminant analysis (LDA) that aims at predicting the model

- The predicted model on the Human genetic data
  - a single out-of-Africa colonization event
  - with admixture to explain Afro-Americans

- With the second forest, the posterior error knowing the observed data ≈ 0.002
- Hence posterior probability of the MAP ≈ 0.998

+

- much faster algorithm than nn methods or local linear methods on a large set of simulations

On the Star Formation History on $\approx 4 \times 10^4$ galaxies

Differences with the population genetic example

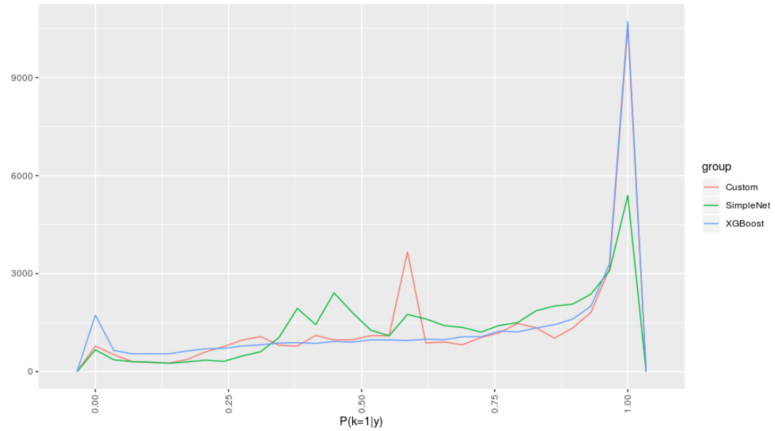
- one observation $\rightarrow \approx 4 \times 10^4$ observed galaxies

$\Longrightarrow$ $\approx 4 \times 10^4$ posterior probabilities to compute

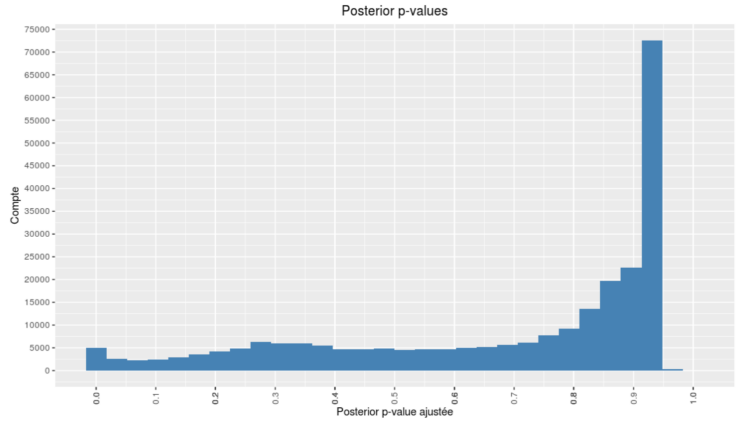
- Once a machine learning method is trained on the simulations, computing the posterior probability of each model is relatively fast

$+$

- Two random forests $\rightarrow$ one **Gradient Boosting Machine**
- We have also computed an ABC posterior p-value of the most complex model



Dist. of posterior probabilities of the most complex model



Distr. of the ABC posterior p-values of the most complex model on all observed galaxies

Conclusion

- When no other inference method at our disposal, ABC is a valuable tool
- ABC is based on a set of simulations from the model(s)
- Replace the dataset $\mathbf{x}$ with $S(\mathbf{x})$, hence replace $p(\cdot|\mathbf{x})$ with $p(\cdot|S(\mathbf{x}))$
- Should be seen as a learning problem on the set of simulations
- Take care of what we want to learn and the learning method
- A package `abcrf` on CRAN which implements our Random Forest methodology