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

Table of Contents

1 Introduction

2 The Machine Learning perspective on ABC

3 Numerical results
1 Introduction

2 The Machine Learning perspective on ABC

3 Numerical results
Bayesian model choice

The evidence of model $M$

- $p(x|M) = \int \pi(\theta|M) f(P|\theta,M) f(x|P,\theta,M) dP d\theta$

where:
- $P$ is the past history
- $x$ is data collected at present
- $M$ is the model index
- $\theta$ the parameters

The MAP model

- selects the model with maximum a posteriori probability

How to compute the posterior probability?

or the MAP model?

Prior/posterior on the collection of models

- prior probability of model $M$:
  $p(M)$
- posterior probability of model $M$:
  $p(M|x) \propto p(M)p(x|M)$
Approximate Bayesian computation (ABC)

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

\[ f(x|\theta) \]

- 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
\[ \implies \text{intractable likelihoods} \]

**ABC** is a computational technique that only requires being able to sample from the likelihood \( f(x|\theta) \).
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 $\theta$ depends on the model
ABC in Astrophysics

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

\[ f(x|\theta) \]

- 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

\[ \implies \text{ABC speeds up the computation} \]

**Almost not approximative**
- We do not summarise the data: each galaxy is represented by a vector of dimension \( \approx 20 \).
- 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?

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

Data: Spectral Energy Distribution (SED) sampled at a few points (\(\tilde{20}\) datapoints per galaxy)
Table of Contents

1 Introduction

2 The Machine Learning perspective on ABC

3 Numerical results
ABC model choice

Simulation algorithm
For $i$ in $1 : N$
- Draw $M_i$ from prior probability
- Draw $\theta_i$ from prior of the model
- Draw a dataset $x_i$ from $M_i, \theta_i$
EndFor

Summarize datasets
- with a non linear
  $S : \text{data space} \rightarrow \mathbb{R}^d$
- to compare $S(x_{\text{obs}})$ & the
  $S(x_i) = (S_1(x_i), \ldots, S_d(x_i))$’s

Questions?
- What can be said about $M| x_{\text{obs}}$
  with the help of the simulations
  $(M_i, x_i)$ drawn from the joint
distribution?

Rejection ABC
- Choose a threshold $\varepsilon$
- Compute the frequency of each
  model among the simulations that
  satisfy $\|S(x_i) - S(x_{\text{obs}})\| \leq \varepsilon$

How to tune $\varepsilon$?
- Set $\varepsilon$ so that the number of
  accepted simulations is $K_{\text{accepted}}$
  $\implies$ Looks like $K$-nearest neighbor
  method
The reference table of simulations

<table>
<thead>
<tr>
<th>model</th>
<th>θ₁</th>
<th>θ₂</th>
<th>...</th>
<th>s₁</th>
<th>...</th>
<th>sₚ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>S(x₁)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lots of simulations drawn from...

The Machine Learning perspective

The reference table
- A large set of $N$ simulations $(M_i, x_i)$ drawn from the Bayesian model:

\[ p(M)\pi(\theta|\mathcal{M})f(P|\theta,M)f(x|\theta,M) \]

Key point
- Train a machine learning algorithm with the $N$ simulations:
  - Response: the model index $M$,
  - Covariates: the summary statistics $S(x)$.

The goal
- Find the MAP model

\[ \iff \]

- Predict the unknown $M$

- Compute the posterior pr.

\[ p(M = 1|x_{\text{obs}}) \]

\[ \iff \]

- Predict the average response

\[ p(M = 1|x_{\text{obs}}) = \mathbb{E} \left( 1\{M = 1\} \middle| x_{\text{obs}} \right) \]

Pudlo, Marin et al. (2016): Reliable ABC model choice via random forests
The Machine Learning perspective on ABC

The ML perspective (continued)

- If the response is $M$ 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.

**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.

**Basic ABC model choice**

- Select the $k$ closest $S(x_i)$’s to the observed data $S(x)$.
- Predict $M$ as the most frequent model among these selected simulations.
- Return the frequency of each model among these simulations (averaging the indicator vectors).

**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.

$\implies$ all local or nn methods suffer from the curse of dimensionality: $\dim d$ of $S(x)$ 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

\[ \hat{M}(x) \]

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? → Conditional error rate knowing \( x \)
The Machine Learning perspective on ABC

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 $\hat{M}(x_i)$ for each simulation
- set $Y_i = 1\{\hat{M}(x_i) \neq M_i\}$
EndFor

Proposition

The conditional error rate $= 1- \text{posterior pr of the MAP}$

\[ \mathbb{E}(1\{\hat{M}(x) \neq M\}|S(x)) = 1- \mathbb{P}(\hat{M}(x)|S(x)) \]

Train a second random forest
- to predict $Y_i$ knowing $x_i$
- 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)
Papamakarios G., Murray I (NIPS, 2016): 
Fast $\varepsilon$-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 → Grégoire Aufort

Results in term of prior error rate where comparable ($\approx 10\%$)
Table of Contents

1 Introduction

2 The Machine Learning perspective on ABC

3 Numerical results
On the Human history

<table>
<thead>
<tr>
<th>Method</th>
<th>prior error(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>9.91</td>
</tr>
<tr>
<td>(k)-nn on (S(x))</td>
<td>23.18</td>
</tr>
<tr>
<td>(k)-nn on LDA axes</td>
<td>6.29</td>
</tr>
<tr>
<td>RF</td>
<td>8.84</td>
</tr>
<tr>
<td>RF on (S(x)) &amp; LDA</td>
<td>5.01</td>
</tr>
</tbody>
</table>

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

- Random forest (RF) is the best classifier if we complements the original summary statistics with projections of \(S(x_i)\) 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 \(\approx 0.002\)

- Hence posterior probability of the MAP \(\approx 0.998\)

- 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

$\Rightarrow \approx 4 \times 10^4$ 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 $x$ with $S(x)$, hence replace $p(\cdot|x)$ with $p(\cdot|S(x))$
- 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 abc.rf on CRAN which implements our Random Forest methodology