Reliable ABC Model Choice via Random Forests

Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gauthier, and Christian P. Robert

Anne Durif, Victor Letzelter

September 3, 2015

Outline

Algorithm 1 - ABC Model choice algorithm

(A) Generate a reference table including N_{ref} simulations $(m, S(x))$ from $\pi(m)\pi(\theta|m)f(x|m,\theta)$
(B) Learn from this set to infer about m at $s_0 = S(x_0)$

☞ Step B : Usually made with local methods (like kNN)
☞ Methods that suffer from **the curse of dimensionality**.

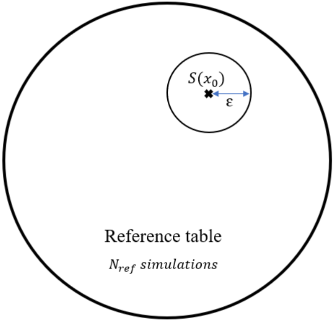

Figure: Inference from a reference table

Algorithm 1 - ABC Model choice algorithm

(A) Generate a reference table including N_{ref} simulations $(m, S(x))$ from $\pi(m)\pi(\theta|m)f(x|m,\theta)$
(B) Learn from this set to infer about m at $s_0 = S(x_0)$

☞ Step B : Usually made with local methods (like kNN)
☞ Methods that suffer from **the curse of dimensionality**.

An example

- Let's consider time series $(X_t^{(i)})_{t \in [\![0,T]\!]}$ whose dynamics are MA(1) or MA(2).
- We only know, as summary statistics, the first N autocorrelations $(\tau_j^{(i)})_{j \in [\![1,N]\!]}$ associated to each time series.
- We want to deduce the model $m^{(i)}$ associated to each time series i.

Random Forests : A New Approach

Why Random Forests ?

- To handle high-dimensional settings.

Figure: Forest Example

Random Forests : A New Approach

Why Random Forests ?

- To handle high-dimensional settings.
- They have a good predictive power

Figure: Forest Example

Random Forests : A New Approach

Why Random Forests ?

- To handle high-dimensional settings.
- They have a good predictive power

☞Vote by majority in classification, averaging in regression

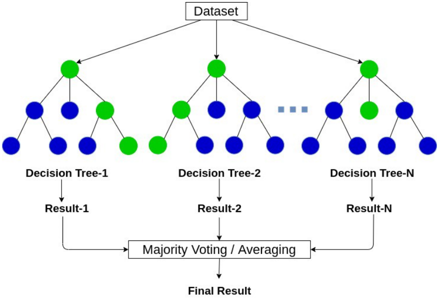

Figure: Random Forest Example

Algorithm 2 - ABC-RF

(A) Generate a reference table including N_{ref} simulation $(m, S(x))$ from $\pi(m)\pi(\theta|m)f(x|m,\theta)$
(B) Construct N_{tree} randomized CART which predict m using $S(x)$
for $b = 1$ **to** N_{tree} **do**
draw a bootstrap (sub-)sample of size N_{boot} from the reference table
grow a randomized CART T_b
end for
(C) Determine the predicted indices for $S(x_0)$ and the trees $T_b; b = 1, ..., N_{tree}$
(D) Affect $S(x_0)$ according to a majority vote among the predicted indexes

Posterior probability estimation

Algorithm 3 - Estimating the posterior probability of the selected model

(A) Computing the binary Error $\mathbf{1}(\hat{m}(s) \neq m)$ for all terms of the test table (with out-of-bag classifiers)

☞ Couples in the form $(s, \mathbf{1}(\hat{m}(s) \neq m))$

(B) Train a random forest to map $s \underset{\rho}{\mapsto} \mathbf{1}(m(s) \neq m)$.

(C) Compute $\rho(S(x_0))$ and return $1 - \rho(S(x_0))$ as our estimate of $\mathbb{P}[m = m(S(x_0))|S(x_0)]$.

Practical example with MA(1) and MA(2)

Def [Moving average]

A stochastic process $(X_t)_{t\in\mathbb{N}}$ in the form $X_t = \mu + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$
Where $\forall i,\ \theta_i \in \mathbb{R}$, and $\varepsilon_t \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$.

Practical example with MA(1) and MA(2)

Def [Moving average]

A stochastic process $(X_t)_{t\in\mathbb{N}}$ in the form $X_t = \mu + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$
Where $\forall i,\ \theta_i \in \mathbb{R}$, and $\varepsilon_t \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$.

Def [Identifiability of the MA(2) Model]

The model MA(2) is identifiable if

$$-2 < \theta_1 < 2, \quad \theta_1 + \theta_2 > -1, \quad \theta_1 - \theta_2 < 1$$

Process

1. Generate a reference table including $N = 10^4$ times series of length $T = 100$, $(m, S(x))$ from $\pi(m)\pi(\theta|m)f(x|m,\theta)$.
   Where θ values are chosen in the cone of identifiability of the models $m \in \{1, 2\}$ and $S(x) = (\tau_i(x))_{i \in [\![1,7]\!]}$.

Process

1. Generate a reference table including $N = 10^4$ times series of length $T = 100$, $(m, S(x))$ from $\pi(m)\pi(\theta|m)f(x|m,\theta)$.
   Where θ values are chosen in the cone of identifiability of the models $m \in \{1, 2\}$ and $S(x) = (\tau_i(x))_{i \in [\![1,7]\!]}$.
2. Train a model $\hat{m}$ to infer m from $S(x) := s$ using $\hat{m}(s)$.

Process

1. Generate a reference table including $N = 10^4$ times series of length $T = 100$, $(m, S(x))$ from $\pi(m)\pi(\theta|m)f(x|m,\theta)$.
   Where θ values are chosen in the cone of identifiability of the models $m \in \{1,2\}$ and $S(x) = (\tau_i(x))_{i \in [\![1,7]\!]}$.
2. Train a model $\hat{m}$ to infer m from $S(x) := s$ using $\hat{m}(s)$.
3. Estimate the posterior probability of the selected model.

Confusion tables for the models

$$N_{ref} = 10^4$$

Random Forest ($\sim 7\ s$)		
True vs. Pred.	1	2
1	4622	380
2	390	4608

☞ Error rate : 7.7%.

KNN with $k = 100$ ($\sim 10\ h$)		
True vs. Pred.	1	2
1	4857	2326
2	742	2075

☞ Error rate : 30.7%.

Logistic regression ($\sim 10\ s$)		
True vs. Pred.	1	2
1	4930	86
2	82	4902

☞ Error rate : 1.7%.

KNN with $k = 50$ ($\sim 7\ h$)		
True vs. Pred.	1	2
1	5016	2970
2	528	1486

☞ Error rate : 35.0%.

Figure: Discrepancy of Posterior Probabilities

How to select summary statistics that are informative for model choice ?

Figure: Discrepancy of Posterior Probabilities

How to select summary statistics that are informative for model choice ?
☞Projection techniques (LDA, Bayes Factor)

Figure: Discrepancy of Posterior Probabilities

How to select summary statistics that are informative for model choice ?
☞Projection techniques (LDA, Bayes Factor)

Choice between MA(1) and MA(2) : Using the first AC ?

Figure: Discrepancy of Posterior Probabilities

How to select summary statistics that are informative for model choice ?
☞Projection techniques (LDA, Bayes Factor)

Choice between MA(1) and MA(2) : Using the first AC ?

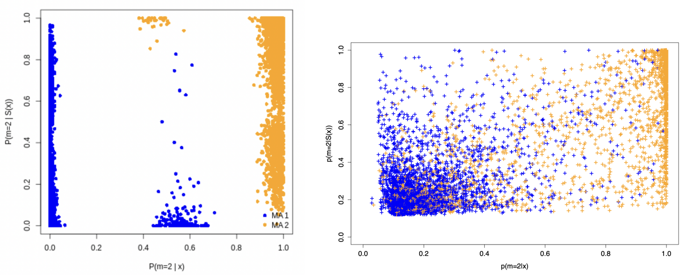

Figure: Discrepancy between posterior probabilities based on the whole data and based on a summary

Figure: Discrepancy of Posterior Probabilities

How to select summary statistics that are informative for model choice ?
☞Projection techniques (LDA, Bayes Factor)

Choice between MA(1) and MA(2) : Using the first AC ?

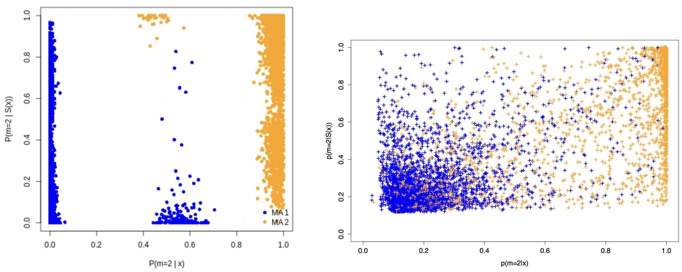

Figure: Discrepancy between posterior probabilities based on the whole data and based on a summary

Conclusion

Advantages of Random Forests

- Only 3 parameters to calibrate : $N_{\text{tree}}, N_{\text{boot}}$ and the feature-selection criterion (Gini index here).
- Lower prior error rate (more accurate).
- Better time complexity (more than 50 times faster than standard ABC).

☞Interview with P. Pudlo.

Conclusion

Advantages of Random Forests

- Only 3 parameters to calibrate : $N_{\text{tree}}, N_{\text{boot}}$ and the feature-selection criterion (Gini index here).
- Lower prior error rate (more accurate).
- Better time complexity (more than 50 times faster than standard ABC).

☞ Interview with P. Pudlo.

Thanks for your attention!

References

- P. Pudlo, J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier, C. P. Robert (2015), Reliable ABC Model Choice via Random Forests.
- S. Allassonnière, *Computational Statistics*, Université de Paris
- L. Breiman (2001), Random Forests. *Machine Learning*, 45, 5–32.
- B. Lakshminarayanan, D.M. Roy, Y.W. Teh (2014), Mondrian Forests: Efficient Online Random Forests.

Applications: SNP and Microsatellite Data

Single Nucleotide Polymorphism Data

- Dataset including 50,000 SNP markers genotyped in four populations: Yoruba (Africa), Han (East Asia), British and American individuals of African ancestry
- Six possible scenarios
- cf. The 1000 Genomes Project Consortium, 2012.

Microsatellite Data

- Invasion routes of the Harlequin ladybird
- Samples from three natural and two biocontrol populations genotyped at 18 microsatellite markers
- Ten possible scenarios

Applications: SNP and Microsatellite Data

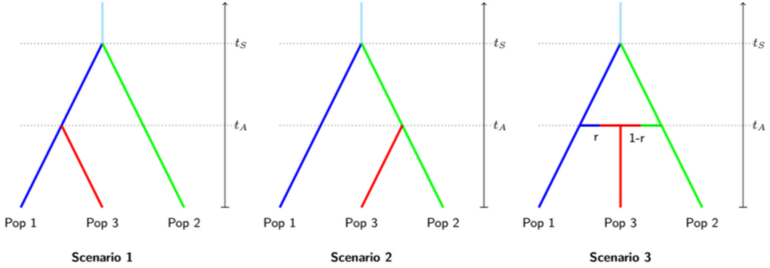

Figure: Population Genetics History: Model Examples

Mondrian Forests

Mondrian Processes

Families of random, hierarchical, binary partitions and probability distributions over tree data structures

Mondrian Trees

- Every node r has a split time τ_r
- τ_r increases with the depth of the tree (increase sampled stochastically)

☞ $\tau_{root} = 0$ and $\tau_{leaves} = \infty$

Sources:

- D. Roy and Y.W. Teh (2008), *The Mondrian Process*
- Data and Knowledge Modeling and Analysis, *Lect 4B - Extra Trees, Gradient Boosting and Hoeffding Trees*