Approximate Bayesian Computation for Model Selection

Yuxi Jiang Natalia Garcia Martin Lorenzo Pacchiardi

8 February 2019

Contents

Motivation

- In a Bayesian setting, we want to compute posterior, i.e. $\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto l(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta})$.
- What if $l(\boldsymbol{\theta}|\boldsymbol{y})$ is not available or expensive?
- Approximate Bayesian Computation offers a solution when we can simulate from prior and likelihood.

Approximate Bayesian Computation

Use a rejection sampling procedure, generating $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and $\boldsymbol{z} \sim f(\boldsymbol{z}, \boldsymbol{\theta})$.

Algorithm 1 ABC algorithm

for $i = 1$ to N **do**
  repeat
    Generate $\boldsymbol{\theta}'$ from the prior distribution $\pi(\cdot)$
    Generate $\boldsymbol{z}$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$
  until $\rho\{\eta(\boldsymbol{z}), \eta(\boldsymbol{y})\} \leq \epsilon$
  Set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$
end for

NB: the choice of threshold ϵ and summary statistics $\boldsymbol{\eta}$ are important.

Approximate Bayesian Computation for Model Choice (ABC-MC)

- Generalization to the case of M models, with model index $\mathcal{M} = 1, 2, \ldots, M$.
- Model $\mathcal{M} = m \to$ parameters $\boldsymbol{\theta}_m$, prior $\pi_m(\cdot)$ and likelihood $f_m(\cdot, \boldsymbol{\theta}_m)$

Approximate Bayesian Computation for Model Choice (ABC-MC)

- Generalization to the case of M models, with model index $\mathcal{M} = 1, 2, \ldots, M$.
- Model $\mathcal{M} = m \rightarrow$ parameters $\boldsymbol{\theta}_m$, prior $\pi_m(\cdot)$ and likelihood $f_m(\cdot, \boldsymbol{\theta}_m)$

Algorithm 3 ABC-MC

for $i = 1$ to N **do**
  repeat
    Generate m from the prior $\pi(\mathcal{M} = m)$
    Generate $\boldsymbol{\theta}_m$ from the prior $\pi_m(\cdot)$
    Generate $\boldsymbol{z}$ from the model $f_m(\cdot|\boldsymbol{\theta}_m)$
  until $\rho\{\boldsymbol{\eta}(\boldsymbol{z}), \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon$
  Set $m^{(i)} = m$ and $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_m$
end for

Estimate of the Bayes Factor

From the output of the previous algorithm, you can estimate the Bayes factor as:

$$\hat{B}_{12}(\boldsymbol{y}) = \frac{\sum_{i=1}^{N} \mathbb{I}_{m^{(i)}=1}}{\sum_{i=1}^{N} \mathbb{I}_{m^{(i)}=2}} \xrightarrow{N\to\infty,\ \epsilon\to 0} B_{12}^{\eta}(\boldsymbol{y}) = \frac{P(\eta(\boldsymbol{y})|\mathcal{M}=1)}{P(\eta(\boldsymbol{y})|\mathcal{M}=2)},$$

while the true Bayes Factor is

$$B_{12}(\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\mathcal{M}=1)}{P(\boldsymbol{y}|\mathcal{M}=2)}.$$

They are in general different, even when η is sufficient for the two models.

Sufficient statistics for model choice

Case where η is sufficient for the two models:

- build an embedding model $\mathcal{M} = 0$, under which the others are nested
- any statistic sufficient for it will be sufficient for model selection, i.e.

$$B_{12}^{\eta}(\boldsymbol{y}) = B_{12}(\boldsymbol{y})$$

Exponential family

Consider models from exponential family:

$$f_i(\boldsymbol{y}|\boldsymbol{\theta}_i) \propto \exp(\eta_i(\boldsymbol{y}) \cdot \boldsymbol{\theta}_i + \gamma_i(\boldsymbol{y})).$$

The following is an embedding model:

$$f_0(\boldsymbol{y}|\boldsymbol{\theta}_0) \propto \exp(\eta_1(\boldsymbol{y}) \cdot \boldsymbol{\theta}_1 + \eta_2(\boldsymbol{y}) \cdot \boldsymbol{\theta}_2 + \alpha_1\gamma_1(\boldsymbol{y}) + \alpha_2\gamma_2(\boldsymbol{y})), \quad \alpha_i \in \{0,1\}.$$

$\Rightarrow \boldsymbol{\eta} = (\eta_1, \eta_2, \gamma_1, \gamma_2)$ is sufficient for model selection

Toy example

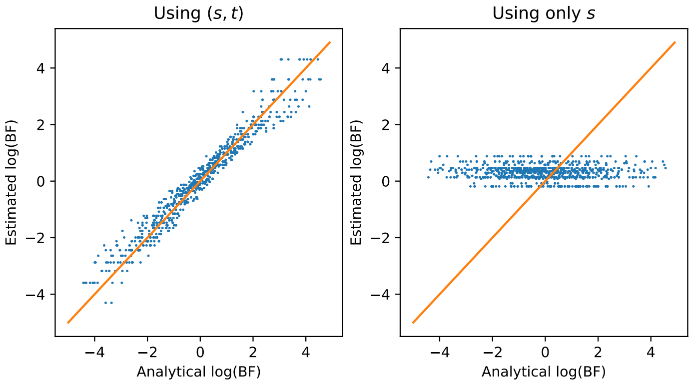
- Model 1: $y_i \sim \text{Poisson}(\lambda)$ iid, $\lambda \sim \text{Exponential}(1)$
- Model 2: $y_i \sim \text{Geometric}(\mu)$ iid, $\mu \sim \text{Uniform}(0,1)$

$\boldsymbol{\eta}(\boldsymbol{y}) = (s,t) = \left(\sum_j y_j, \sum_j \log y_j!\right)$ is a sufficient statistic for model selection. Note that s alone is a sufficient statistic for each model, but not for model selection.

Toy example

- Model 1: $y_i \sim \text{Poisson}(\lambda)$ iid, $\lambda \sim \text{Exponential}(1)$
- Model 2: $y_i \sim \text{Geometric}(\mu)$ iid, $\mu \sim \text{Uniform}(0, 1)$

$\boldsymbol{\eta}(\boldsymbol{y}) = (s, t) = \left(\sum_j y_j, \sum_j \log y_j!\right)$ is a sufficient statistic for model selection. Note that s alone is a sufficient statistic for each model, but not for model selection.



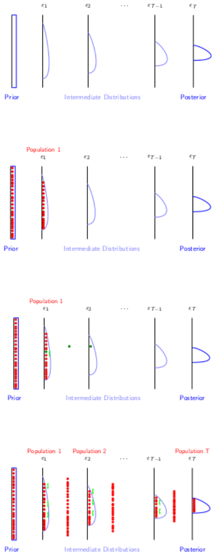
General considerations

- Without sufficient statistics for the two models, previous analysis is not applicable.
- Condition for the approximate Bayes Factor to be asymptotically convergent[1]: expectation of the summary statistic to differ asymptotically under the two models.
- Ideal sufficient statistic is therefore an ancillary one with different expectation under the two models.

[1] [Marin *et al.* , 2014]

ABC Sequential Monte Carlo [Toni *et al.* , 2008]

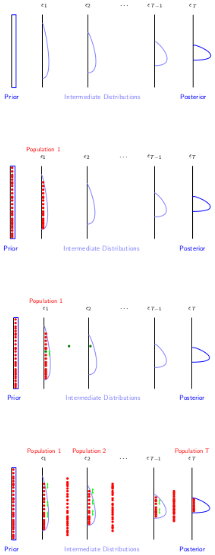
- Derived from Sequential Importance Sampling (SIS)
- Define a tolerance set $\{\epsilon_0, \ldots, \epsilon_T\}$ such that $\epsilon_0 > \cdots > \epsilon_T \geq 0$
- Construct intermediate distributions $\pi(\boldsymbol{\theta} | d(x^*, y) \leq \epsilon_t)$, $t = 0, \ldots, T-1$ which converges to the posterior distribution $\pi(\boldsymbol{\theta} | d(x^*, y) \leq \epsilon_T)$



[1][Toni & Stumpf, 2009]

ABC Sequential Monte Carlo [Toni *et al.* , 2008]

- For $t = 0$, sample directly from the prior distribution to obtain N accepted particles and attach a weight to each of particles in the population
  - Population 0: $\{\theta_0^{(1)}, \ldots, \theta_0^{(N)}\}$
- For $0 < t \leq T$, sample from the previous population with weights and perturb the particle using a perturbation kernel
  - Population t: $\{\theta_t^{(1)}, \ldots, \theta_t^{(N)}\}$
- Population T is a sample from the approximated posterior distribution



[1] [Toni & Stumpf, 2009]

ABC-SMC for Model Selection

- Generalize to model selection with M models, $m \in \{1, \ldots, M\}$
- Model-specific parameters $\theta(m) = (\theta(m)^{(1)}, \ldots, \theta(m)^{(k_m)})$
- Perturbation kernel K_t

ABC-SMC for Model Selection

Algorithm 4 ABC-SMC algorithm for model selection

1. Initialize $\epsilon_0, \ldots, \epsilon_T$, set $t = 0$, $i = 1$.
2. Sample m^* from $\pi(m)$
   - If $t = 0$, sample θ^{**} from $\pi(\theta(m^*))$.
   - If $t > 0$, sample θ^* from the previous population $\{\theta(m^*)_{t-1}\}$ with weights $w(m^*)_{t-1}$, perturb the particle θ^* to obtain $\theta^{**} \sim K_t(\theta|\theta^*)$.
3. If $\pi(\theta^{**}) = 0$, go back to Step 2
4. Simulate a candidate dataset $x^* \sim f(x|\theta^{**}, m^*)$, if $d(x^*, y) \geq \epsilon_t$, go back to Step 2
5. Set $m_t^{(i)} = m^*$, add θ^{**} to the population of particles $\{\theta(m^*)_t\}$, and calculate its weight as

$$w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0 \\ \dfrac{\pi(\theta^{**})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta^{**})}, & \text{if } t > 0 \end{cases}$$

6. If $i < N$, set $i = i + 1$ and go back to Step 2.
7. Normalise the weights for every m
8. If $t < T$, set $t = t + 1$ and go back to Step 2

SIR Model Selection

- Model 1

$$\dot{S} = \alpha - \gamma SI - dS$$
$$\dot{I} = \gamma SI - \nu I - dI$$
$$\dot{R} = \nu I - dR$$

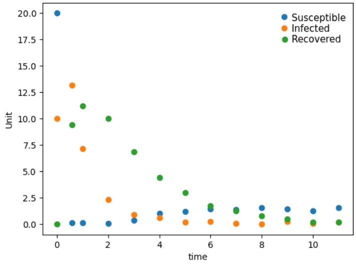
- Model 2

$$\dot{S} = \alpha - \gamma SI - dS$$
$$\dot{L} = \gamma SI - \delta L - dL$$
$$\dot{I} = \delta L - \nu I - dI$$
$$\dot{R} = \nu I - dR$$

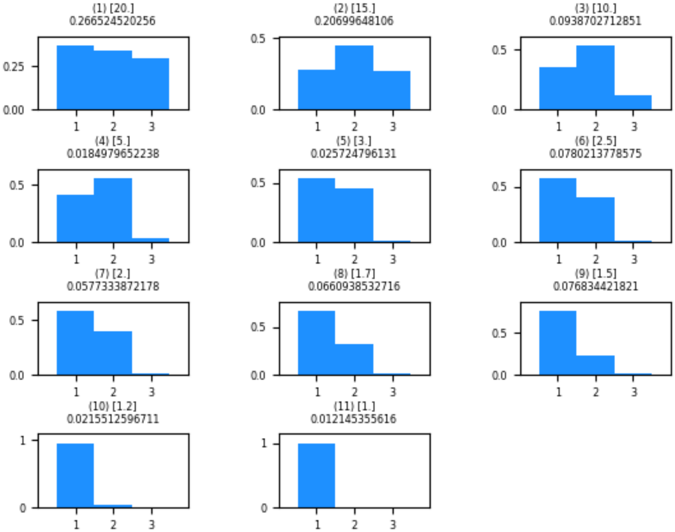
- Model 3

$$\dot{S} = \alpha - \gamma SI - dS + eR$$
$$\dot{I} = \gamma SI - \nu I - dI$$
$$\dot{R} = \nu I - dR - eR$$

- Data simulated using Model 1
- 1000 particles



SIR Model Selection



ABC model choice via random forests (ABC-RF)

Algorithm 5 ABC-RF [Pudlo *et al.* , 2015]

1. Generate a reference table made of the set of $(m,\ \boldsymbol{\eta}_m(\boldsymbol{z}))$ from the N simulations using $\pi(\mathcal{M}=m)$, $\pi_m(\boldsymbol{\theta}_m)$ and $f_m(\boldsymbol{z}|\boldsymbol{\theta}_m)$
2. Construct N_{tree} randomised CART which predict m using $\boldsymbol{\eta}_m(\boldsymbol{z})$

    for $b=1$ to N_{tree} **do**

        draw a bootsrap (sub-)sample of size N_{boot} from the reference table

        grow a randomised CART T_b

        for $n=1$ to N_{nodes} **do**

            Select n of the predictors at random

            Determine the best split from among those predictors

        end for

    end for

3. Determine the predicted indexes for $\boldsymbol{\eta}(\boldsymbol{y})$
4. Determine $\hat{m}$ according to a majority vote across the predicted indexes

ABC-RF

Algorithm 6 Estimating the posterior probability of the selected model

1. Use the trained RF (Algorithm 5) to predict model by $\hat{m}(\boldsymbol{\eta}(\boldsymbol{z}))$ for each $(m,\ \boldsymbol{\eta}_m(\boldsymbol{z}))$ in the reference table and compute the out-of-bag classifier error $\mathbb{I}(\hat{m}(\boldsymbol{\eta}) \neq m)$
2. Use the reference table to build a RF regression function $\rho(\boldsymbol{\eta})$ regressing the model prediction error $\mathbb{I}(\hat{m}(\boldsymbol{\eta}) \neq m)$ on the summary statistics. $\rho(\boldsymbol{\eta})$ is an estimate of $\mathcal{P}\left[m \neq \hat{m}(\boldsymbol{\eta})|\boldsymbol{\eta}\right]$
3. Apply the RF to the actual observations summarised as $\boldsymbol{\eta}(\boldsymbol{y})$ and return $1 - \rho(\boldsymbol{\eta}(\boldsymbol{y}))$ as the estimate of $\mathcal{P}\left[m = \hat{m}(\boldsymbol{\eta}(\boldsymbol{y}))|\boldsymbol{\eta}(\boldsymbol{y})\right]$

Model selection for SNP data using ABC-RF

- 1,000 autosomal SNP (single nucleotide polymorphisms) markers
- Reference table of 10,000 simulations
- 48 DIYABC summary statistics

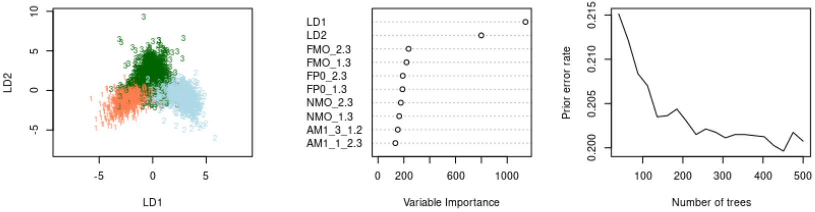


Figure: (Left) LDA projection of reference table. (Middle) Top 10 statistics. (Right) Prior error rate versus number of trees for the ABC-RF model.

Conclusion

- Conditions for the choice of summary statistics in order to get coherent estimates of the Bayes factor
- Advantages of ABC-SMC
  1. Computationally efficient
  2. Higher acceptance rate
- Advantages of ABC-RF
  1. Larger discriminative power
  2. More robust
  3. Computationally efficient
  4. Approximation of posterior probability

References

Marin, Jean-Michel, Pillai, Natesh S, Robert, Christian P, & Rousseau, Judith. 2014.

Relevant statistics for Bayesian model choice.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), **76**(5), 833–859.

Pudlo, Pierre, Marin, Jean-Michel, Estoup, Arnaud, Cornuet, Jean-Marie, Gautier, Mathieu, & Robert, Christian P. 2015.

Reliable ABC model choice via random forests.

Bioinformatics, **32**(6), 859–866.

Toni, Tina, & Stumpf, Michael PH. 2009.

Tutorial on ABC rejection and ABC SMC for parameter estimation and model selection.

arXiv preprint arXiv:0910.4472.

Toni, Tina, Welch, David, Strelkowa, Natalja, Ipsen, Andreas, & Stumpf, Michael PH. 2008.

Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems.

Journal of the Royal Society Interface, **6**(31), 187–202.