Estimating likelihood from simulations in bayesian inference of ecological models with historical contingency

Stéphane DUPAS^{1,2}

1 Laboratoire Evolution, Génome, Comportement et Écologie, Univ Paris-Sud, CNRS, IRD 247, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France

2 Departamento de Ciencias Biologicas, Universidad de Los Andes, Bogota, Colombia

* E-mail: Stephane.Dupas@ird.fr

Abstract

Ecological patterns result from historical contingency and deterministic processes. Taking apart these processes to extract probabilistic models of ecological dynamics is of major importance for ecological forecasting. Due to the high dimensionality of historical contingency it is usually difficult to sample history from observed patterns. In environmental population genetics, the number of possible genealogies linking genetic data and environmental data through demographic and niche models is almost infinite. In ecosystem dynamics time series, the patterns are determined as much by probabilistic model parameters, as by historical variables contingency. Approximate bayesian computation allows to use simulations to approximate this inference process. The rationale is to simulate data, extract summary statistics and retain in the posterior, the parameters values that produced simulations with summary statistic close to observed summary statistics. The major drawbacks of this approach is that summary statistics distance is not exhaustive regarding model likelihood and may bias the results. In the present work, I show that if we can simulate the historical contingency from observed data and probabilistic model in a backward approach, we can use the simulations to estimate a pseudo likelihood that can be used in bayesian inference. I apply to genealogy sampling in environmental demogenetics and ecosystem dynamics modelling.

Introduction

Bayesian inference is a powerful approach to combine prior information to field data and improve models instead of rejecting them. However complex model bayesian sampling can become rapidly intractable, particularly for biological and social processes in which historical contingency plays a major role. Aproximate bayesian computation has become a popular solution for sampling historical processes in ecology [1]. The parameters filtering is obtained through the selection of simulations that are the closest to the observations. However selecting simulations brings back to the problem of defining summary statistics that represent the probabilistic model. Simulation and particle filtering using MCMC have been proposed for dynamic bayesian network [2]. Here I propose to use simulation to complete data in order to estimate an approximate likelihood for bayesian inference. I claim that the best summary statistics would be the probability to observe altogether data and simulations knowing model parameters. In many case this probability can be calculated. The only condition is to be able to formulate a probabilistic model to simulate history from observed data and model. For instance, it is possible to calculate the probability of a genealogy, knowing (1) genetic distributions in allelic and geographic space, (2) environmental data and (3) an environmental demogenetic model including mutation model, niche model and dispersion model. If we average this probability over genealogical simulations we get an estimate proportional to the probability of the parameter set.

Model

Probability and simulation of hidden history knowing indicator data and ecosystem model

We divide the model M into the sampling model M_S that link ecosystem variable history H to data D , and the ecosystem history model M_H that drive ecosystem H variables dynamics ($M = \{M_H, M_S\}$).

Each has its set of parameters θ_S and θ_H . We know how to calculate $p(D/H, M_S)$ and $p(H/M_H)$.

The probability of a given history knowing data and model is given from bayesian rule:

$$p(H/D, M_S, M_H) = p(D/H, M_S, M_H) \frac{p(H, M_H, M_S)}{p(D)} \quad (1)$$

Sampling model M_S and ecosystem model M_H are independent processes : sampling of data from M_S and H , is independent from sampling of history H from ecosystem model M_H . We then have : $p(D/H, M_S, M_H) = p(D/H, M_H)$ and $p(H, M_H, M_S) = p(H, M_H) \times p(M_S)$. Since we have also $p(H, M_H) = p(H/M_H) \times p(M_H)$, equation (1) gives:

$$p(H/D, M) = p(H/M_S, M_H, D) = p(D/H, M_S) \times p(H/M_H) \times \frac{p(M_H) \times p(M_S)}{p(D)} \quad (2)$$

From this distribution, we can simulate a history H_i by using prior information on M_H , M_S , and D . D is independent on model parameters. We also have the probability of each historical simulation H_i knowing data and model using (2).

Estimation of model likelihood by averaging probability of histories

Now we can estimate the probability of the data knowing the model. We consider $[H] = H_i, i \in [1..n]$ the set of all possible histories.

$$\begin{aligned} \sum_{H_i \in [H]} p(H_i/D, M) &= \sum_{H_i \in [H]} p(D/H_i, M_S) \times p(H_i/M_H) \times \frac{p(M_H) \times p(M_S)}{p(D)} \\ &= p(D/M_H, M_S) \times \frac{p(M_H) \times p(M_S)}{p(D)} \\ &= p(M_H, M_S/D) \end{aligned} \quad (3)$$

We see in the equation that the sum of data probability knowing sampling model from history multiplied by history probabilities knowing ecosystem model plays the role of a likelihood function in the classic bayesian equation.

We can have an estimation of the pseudo likelihood $\sum_{H_i \in [H]} p(D/H_i, M_S) \times p(H_i/M_H)$ by simulating a number of times H_i using distribution in (2) and averaging the pseudo likelihood. Since the simulation will produce the most probable histories, after a number of simulations, we will get an estimate proportional to the probability of all histories.

We can therefore name this method bayesian inference with aproximated likelihood (BIWAL).

I now apply to ecosystem and coalescent models.

Algorithm

The following algorithm will be used :

1. Set sampling and ecosystem models $p(D/H, M_S)$ and $p(H/M_H)$
2. Set sampling and ecosystem model parameters priors $p(\theta)$, where $\theta = [\theta_S, \theta_H]$, and θ_S , and θ_H are the M_S and M_H model parameters, respectively.
3. Set Metropolis sampling generation $g = 1$, a parameter proposal rule from θ to θ' , and a *thining* value for sampling frequency of θ
4. Sample θ_g from prior and calculate posterior θ_g the same way as described in steps 6 and 7 for θ'
5. Propose a θ' from θ using metropolis proposal rule
6. Repeat for $i \in 1 : n$
 - (a) Simulate History $H_i \sim p(H_i/M, D)$
 - (b) Calculate $p(H_i/M, D)$
7. calculate posterior $p(\theta'/D)$ from equation (8)
8. if $p(\theta') > p(\theta)$ set $\theta = \theta'$
 if $p(\theta') \leq p(\theta)$ set $\theta = \theta'$ with probability $\frac{p(\theta')}{p(\theta)}$

9. save θ_g if g is a multiple of *thining*

10. $g = g + 1$ go to step 5.

Application to coalescent model

In Dupas (unpublished), we developped an algorithm to simulate and estimate the probability of a genealogy of individuals knowing their spatial and genetic distributions, D and G , respectively, the environmental data E , and the environmental demogenetic model M_D, M_G (equation 13, Dupas 2018). Consider a simulated genealogy $H_i = \text{genealogy}_i$ sampled from $p(\text{genealogy}/M_D, M_G, G, D, E)$ then according to (6)

$$p(M_D, M_G/G, D, E) \propto \lim_{n \rightarrow \infty} \sum_{i=1}^n p(\text{genealogy}_i/M_D, M_G, G, D, E)/n \times p(M_P, M_G) \quad (4)$$

Application to crop yield prediction

We develop here a crop yield model we apply the growth rate function based on existing apriori deterministic model constructed from literature information [3].

1. $D_{nr,i,j}$ = number of rainy days in month i , year j

2. $D_{Pr,i,j}$ = precipitation

3. $D_{T,i,j}$ = average daily temperature

4. $D_{Tx,i,j}$ = maximum daily temperature

5. $D_{Tn,i,j}$ = minimum daily temperature

6. $D_{ETP,i,j}$ = potential evapotranspiration

7. $D_{Y,10,j}$ = observed maize yield at harvest

8. $H_{Y,10,j}$ = expected maize yield at harvest
9. $H_{M,i,j}$ = plant mass
10. $H_{vgr,j}$ = Maize variety maximum growth rate
11. $H_{vprd,j}$ = Maize variety potential resistance to drought (decrease in yield with respect to per unit decrease in ETa) [3]
12. $H_{ETA,i,j}$ = realized evapotranspiration

In addition we consider θ_{Tgr} the temperature coefficient of growth rate. We dispose of monthly data for each of these historical variables, for the 94 French departments for 38 years. In this example there is no missing data.

The ecosystem model is as follows :

$$\begin{aligned}
 H_{M,i} &\rightarrow N(\mu = H_{M,i-1} \times H_{vgr,i} \times \theta_{Tgr} \times H_{T,i} [1 - H_{vprd,j} \frac{H_{ETA,i}}{D_{ETP,i}}], \sigma = \sigma_m) \\
 H_{ETA,i} &\rightarrow N(\mu = D_{ETP,i} - D_{Pr,i,j}, \sigma = \sigma_{ETA}) \\
 H_{vgr,j} &\rightarrow N(\mu = H_{vgr,j-1}^9 - H_{M,10,j-1}, \sigma = \sigma_{vgr}) \\
 H_{vprd,j} &\rightarrow N(\mu = H_{vprd,j}, \sigma = \sigma_{vprd}) \\
 H_{Y,10,j} &\rightarrow N(\mu = D_{Y,10,j}, \sigma = \sigma_Y)
 \end{aligned}$$

Discussion

I propose a new method to sample complex ecological models with important historical contingency. The rationale is to average historical processes calculating simulations probability and to sample parameters using Metropolis Hasting algorithm. The requisite is to be able inverse part of the model using bayesian rules in order to simulate history from data. This backward approach is classic in populations genetics

where coalescent theory has been developed since the 1980's [4]. But similar models can be developed in other area in ecology. In particular for ecosystem dynamics inference from time series data [5].

Acknowledgments

This work has been supported by IRD EGCE research unit project and BASC labex

References

1. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*. 2009;doi:10.1098/rsif.2008.0172.
2. Liu, J and West M. Combining parameter and state estimation in simulation-based filtering. In: Doucet A, de Freitas N, Gordon NJ, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York.; 2001. p. 197–224.
3. Djaman K, Irmak S, Rathje WR, Martin DL, Eisenhauer DE. Maize evapotranspiration, yield production functions, biomass, grain yield, harvest index, and yield response factors under full and limited irrigation. 2013;.
4. Kingman J. On the genealogy of large populations. *Journal of Applied Probability*. 1982;19:27–43. doi:doi:10.2307/3213548.
5. Sander EL, Wootton JT, Allesina S. Ecological Network Inference from Long-Term Presence-Absence Data. *Scientific Reports*. 2017;doi:10.1038/s41598-017-07009-x.