Resampling : an improvement in varying population size models

Coralie Merle, Raphaël Leblois, Pierre Pudlo

Resampling : an improvement in varying population size models

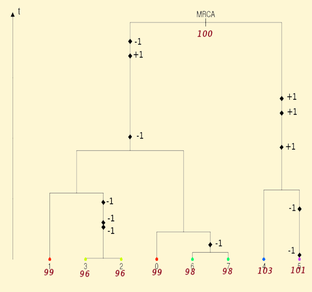
C. Merle[1&2], R Leblois [2] and P. Pudlo[1&2]

1 - Institut de Mathématiques et Modélisation de Montpellier (I3M, UM); 2 - Centre de Biologie pour la gestion des populations (CBGP, INRA).

Genetic polymorphism modelling

Evolution model

- A sample of n gene copies at a single locus from the population of effective size $N(t)$.
- For any given locus, each individual has exactly one ancestor in the previous generation.
- The ancestral relationships between the individuals of the sample going back in time to the MRCA are described by a **gene tree**, distributed according to the n-coalescent.

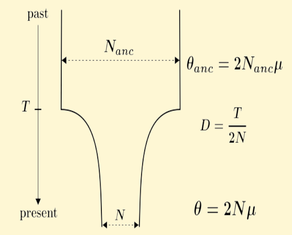


Gene tree of microsatellite markers.

Demographic model

We consider a demographic model, where the effective population size varies in the time, denoted $N(t)$. In particular we work with an **Exponentially Contracting Population** (ECP). If we look backward in time, we have :

$$\tilde{N}(s) = \begin{cases} N\left(\frac{N_{\text{anc}}}{N}\right)^{s/D} & \text{if } 0 \leq s \leq D \\ N_{\text{anc}} & \text{if } s \geq D. \end{cases}$$



Exponentially contracting population model.

Likelihood of the data

Notations: $\mathbf{n}_{\text{obs}}$: observed data, $\mathcal{H}_{\mathbf{n}_{\text{obs}}}$: set of compatible histories with the observed data,

Problem

The histories are not observed. The likelihood of the data is obtained by **summing over all the possibilities** (not observed), which is not feasible in practice:

$$L(\mathbf{n}_{\text{obs}}|\phi) = \int \mathbf{1}_{\{h \in \mathcal{H}_{\mathbf{n}_{\text{obs}}}\}} P_\phi(h) \mathrm{d}h$$

Solution

A class of Monte-Carlo methods, based on Sequential Important Sampling (SIS), allows the likelihood calculation despite the hidden process. The efficiency of these methods was been proven by [1], [2], [3] and [6].

In this schema, the IS distribution propose histories which contribute the most to the sum defining the likelihood. But **changing effective population size introduce a strong inhomogeneity** in the Wright-Fisher model and **the IS distributions become inefficient**. The computation time strongly increases for the same accuracy of the likelihood estimation, so that we can not have a correct estimation.

Evaluating the likelihood with SIS

$$L(\mathbf{n}_{\text{obs}}|\phi) = \int \mathbf{1}_{\{h \in \mathcal{H}_{\mathbf{n}_{\text{obs}}}\}} \frac{P_\phi(h)}{Q_\phi(h)} Q_\phi(h) \mathrm{d}h = \mathbb{E}_{H \sim Q_\phi}\left[\mathbf{1}_{\{H \in \mathcal{H}_{\mathbf{n}_{\text{obs}}}\}} \frac{P_\phi(H)}{Q_\phi(H)}\right]$$

We approximate the likelihood with a Monte-Carlo estimator:

$$\widehat{L(\mathbf{n}_{\text{obs}}|\phi)} = \frac{1}{n_H} \sum_{j=1}^{n_H} \mathbf{1}_{\{H^{(j)} \in \mathcal{H}_{\mathbf{n}_{\text{obs}}}\}} \frac{P_\phi(H^{(j)}|\phi)}{Q_\phi(H^{(j)})} = \frac{1}{n_H} \sum_{j=1}^{n_H} w^{(j)},$$

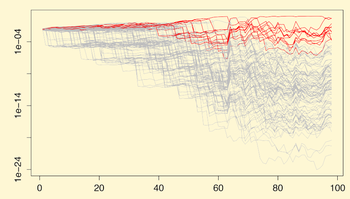
where $H^{(j)} \sim Q_\phi$ are n_H realizations of the latent inhomogeneous Markov process, Q_ϕ is the important sampling distribution and $w^{(j)}$ are the importance weights computed sequentially as follow:

$$w_k^{(j)} = \prod_{i=1}^{k} \frac{P_\phi(H_{i-1}^{(j)}, H_i^{(j)}|\phi)}{Q_\phi(H_i^{(j)}, H_{i-1}^{(j)})} = w_{k-1}^{(j)} \frac{P_\phi(H_k^{(j)}|H_{k-1}^{(j)}, \phi)}{Q_\phi(H_{k-1}^{(j)}|H_k^{(j)})}.$$

Correction of Importance Sampling distribution by Resampling (SISR)

The idea is to resample during the backward building of the histories, so that we learn which are the histories proposed by the IS distribution that really contribute most of the sum defining the likelihood and so save computation time.

We stop the SIS algorithm that builds the histories in parallel at a given checkpoint and we **draw a new collection of n_H histories** according to a **multinomial distribution** with parameter n_H and a **resampling probability distribution v** that could be any distribution. This new algorithm is called **SISR**, see [5].

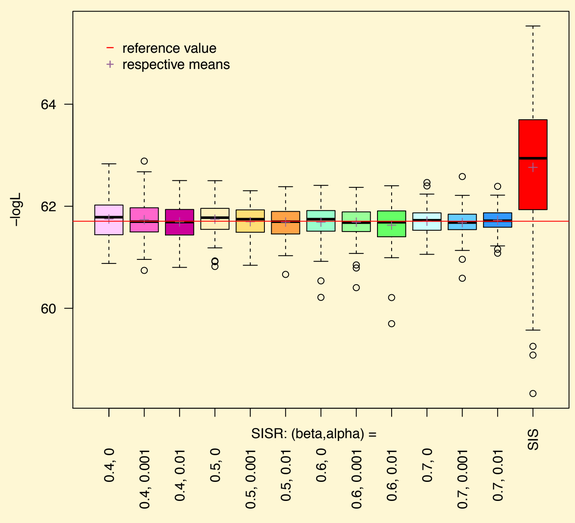


In this paper, we propose a new resampling probability distribution **v** on the collection of histories. Since the information is mainly in the dependency between individuals of the sample, we propose to substitute the **pairwise composite likelihood**, noted $L_2(h|\phi)$, for the likelihood of the end of the history. It does not approximate the likelihood but it behaves the same way and it is fast to compute. Indeed, it is the product of the likelihoods of each pair of individuals in the sample. We propose to draw the j-st history $h^{(j)}$ with probability :

$$v^{(j)} = \left(w^{(j)}\right)^\alpha \left(L_2(h^{(j)}|\phi)\right)^\beta, \ \alpha, \beta \in [0,1], \beta \ll 1.$$

Numerical results when comparing SIS and SISR

With the same computational cost, the Mean Square Error (MSE) between the true value of the likelihood and its estimate is lower when estimating with resampling (SISR) than without (SIS), for a range of values of α and β : $0.4 \leq \alpha \leq 0.7,\ 0 \leq \beta \leq 0.01$.



Boxplots of 100 estimates of the likelihood in a given parameter point with different estimation algorithms, on simulated data sets.

The parameter of interest is the vector $(\theta, D, \theta_{anc})$. **We estimate it by maximum likelihood inference** when the likelihood of the data is estimated by the SIS or SISR algorithm. In all the following cases, the inference is difficult because of the strengh of the contraction or because it happened recently (see [4]).

With the same computational cost, **the resampling allows to reduce the bias and MSE** in the parameter estimation.

$(\theta, D, \theta_{anc})$	=	(0.4,	1.25, 400)	(0.4,	0.25, 40)	(0.4,	0.25, 400)
Algorithm		SIS	SISR	SIS	SISR	SIS	SISR
Rel. bias	θ	0.56	0.34	1.07	0.23	4.9	1.1
	D	−0.02	−0.017	0.23	0.571	−0.06	0.15
	θ_{anc}	0.048	−0.042	0.032	0.023	0.044	−0.053
Rel. RMSE	θ	0.71	0.53	2	1.8	5.2	1.7
	D	0.14	0.14	0.45	0.8	0.14	0.37
	θ_{anc}	0.37	0.38	0.29	0.27	0.25	0.23

Comparison of relative bias and Root Mean Squar Error (RMSE), analysis of data sets simulated under the ECP model, by SIS and SISR with $n_H = 100$ (left) or $n_H = 2000$ (center and right) histories sampled.

References

[1] Maria De Iorio and Robert C Griffiths. Importance sampling on coalescent histories. i. *Advances in Applied Probability*, pages 417–433, 2004.
[2] Maria De Iorio and Robert C Griffiths. Importance sampling on coalescent histories. ii: Subdivided population models. *Advances in Applied Probability*, 36(2):434–454, 2004.
[3] Maria De Iorio, Robert C Griffiths, Raphael Leblois, and François Rousset. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical population biology*, 68(1):41–53, 2005.
[4] Raphaël Leblois, Pierre Pudlo, Joseph Néron, François Bertaux, Champak Reddy Beeravolu, Renaud Vitalis, and François Rousset. Maximum likelihood inference of population size contractions from microsatellite data. *Molecular biology and evolution*, page msu212, 2014.
[5] Jun S Liu, Rong Chen, and Tanya Logvinenko. A theoretical framework for sequential importance sampling with resampling. In Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors, *Sequential Monte Carlo methods in practice*, pages 225–246. Springer, 2001.
[6] Matthew Stephens and Peter Donnelly. Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635, 2000.

Acknowledgments

Labex NUMEV