Acceleration of important sampling methods for the calculation of likelihood in population genetics

Coralie Merle, Raphaël Leblois, J.-M Marin, P Pudlo, F. Rousset

Acceleration of important sampling methods for the calculation of likelihood in population genetics.

C. Merle[1&2], R Leblois [2], J.-M. Marin[1], P. Pudlo[1&2] and F. Rousset[3]
1 - Institut de Mathématiques et modÉlisation de Montpellier (I3M); 2 - Centre de Biologie pour la gestion des populations (INRA); 3 - Institut des Sciences de l'Évolution de Montpellier (ISEM)

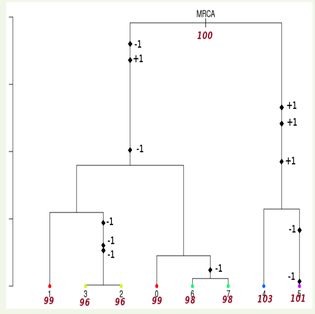
Abstract

- **Model**: The population evolves under a Wright-Fisher model. Hence, the sample evolves according to the Kingman coalescent.
- **Problem**: The likelihood is the sum over all possible histories (not observed), which is not feasible in practice.
- **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 [5].
In the IS sheme, the importance sampling distributions propose histories which contribute most to the sum. But these distribution are not efficient for equilibrium population models and the computation time strongly increases for the same accuracy of the likelihood estimation, so that we can not have a correct estimation.
- **Improvement** : For changing population size model, we decide to use : Sequential Important Sampling with Resampling (SISR). The idea is to resample, during the backward building of the histories, so that we learn wich are the histories proposed by the IS distribution which really contribute most of the sum and so save computation time.

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.

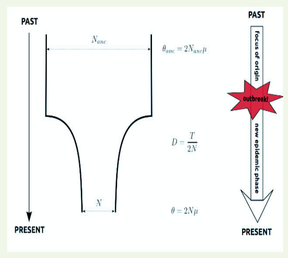


Gene tree of microsatellite markers.

Demographic model

We consider a demographic model, never treated before, where the population effective size varies in the time, notes $N(t)$. In particular we work with an **Exponentially Contracting Population**. If we look backward in time, we have :

$$N(t) = \begin{cases} N_0 \left(\frac{N_{\text{anc}}}{N_0}\right)^{t/D} & \text{si } 0 \leq t \leq D \\ N_{\text{anc}} & \text{si } t \geq D. \end{cases}$$



Exponentially contracting population model.

Likelihood of the data

The histories are not observed. The likelihood of the data is obtained by **summing over all the possibilities** :

$$\begin{aligned} \text{Prob}(\mathbf{n}_{\text{obs}}|\theta) &= \int_{\mathcal{H}} p_0(\mathbf{n}_0) \prod_{\ell=1}^{m+1} \left(p_{s_\ell}(\mathbf{n}_\ell|\mathbf{n}_{\ell-1}) f(s_\ell|\mathbf{n}_{\ell-1}, s_{\ell-1}) \right) \mathrm{d}H \\ &= \int g(\mathbf{n}_{\text{obs}}, H|\theta) \mathrm{d}H. \end{aligned}$$

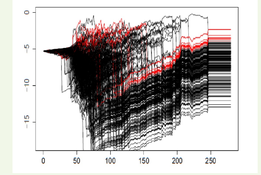
Where :
$\mathbf{n}_{\text{obs}}$: observed data,
$(\mathbf{n}_0, \ldots, \mathbf{n}_{m+1})$ count vector of lenght $(m+2)$, such as $\mathbf{n}_m = \mathbf{n}_{\text{obs}}$ and $|\mathbf{n}_{m+1}| = |\mathbf{n}_{\text{obs}}| + 1$,
$s_0, s_1, s_2, \ldots$: dates of jump (in forward time),
$g(\mathbf{n}_{\text{obs}}, H|\theta) = \mathbf{1}\{H \in \mathcal{H}\} p(H)$,
$\mathcal{H}$: set of compatible histories with the observed data.

Correction of Importance Sampling distribution by resampling (SISR)

Changing effective population size introduce a strong inhomogeneity in the WF model and the IS distributions become inefficient.
We decide to **resample** in our collection of simulated histories :

- to prune the bad histories,
- to produce multiple copies of good histories, to generate futur better histories.

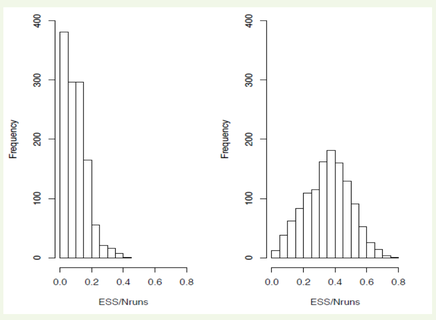
How ? We stop the SIS algorithm that builds the genealogies in parallel at a given time and we modify the composition of the histories collection according to the partial importance weights at this date. This new algorithm is called **SISR**, voir [4].



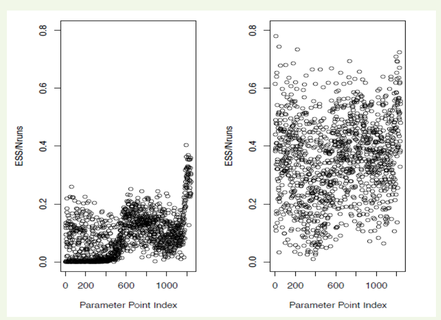
Numerical results when comparing SIS and SISR

Our parameter of interest is the vector $(\theta, D, \theta_{anc})$. We try to estimate this parameter by maximum likelihood inference. The likelihood of the data is estimated by the SIS or SISR algorithm.

Comparison of relative Effective Sample Size when the true parameter is $\theta = 0.4$, $D = 0.25$ and $\theta_{anc} = 40$.



Histogram



Plot

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

		SIS	SISR
Rel. bias	θ	0.56	0.364
	D	−0.0201	−0.0308
	θ_{anc}	0.0479	−0.138
RMSE	θ	0.711	0.557
	D	0.142	0.142
	θ_{anc}	0.369	0.305

With $\theta = 0.4$, $D = 1.25$ and $\theta_{anc} = 400$.

		SIS	SISR
Rel. bias	θ	4.92	1.62
	D	−0.0606	0.177
	θ_{anc}	0.0438	−0.00967
RMSE	θ	5.17	1.82
	D	0.141	0.417
	θ_{anc}	0.245	0.21

With $\theta = 0.4$, $D = 0.25$ and $\theta_{anc} = 400$.

		SIS	SISR
Rel. bias	θ	1.35	0.188
	D	0.191	0.687
	θ_{anc}	0.0522	0.0196
RMSE	θ	2.98	1.82
	D	0.442	0.909
	θ_{anc}	0.322	0.267

With $\theta = 0.4$, $D = 0.25$ and $\theta_{anc} = 40$.

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] 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.
[5] 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