Acceleration of important sampling methods for the calculation of likelihood in population genetics
Coralie Merle, Raphaël Leblois, J.-M Marin, P Pudlo, F. Rousset

To cite this version:
Coralie Merle, Raphaël Leblois, J.-M Marin, P Pudlo, F. Rousset. Acceleration of important sampling methods for the calculation of likelihood in population genetics. Day of The Institute for Computational Biology (IBC), May 2014, Montpellier, France. 2014. hal-02932305

HAL Id: hal-02932305
https://hal.inrae.fr/hal-02932305
Submitted on 7 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 proven by [1], [2], [3] and [5]. In the IS scheme, the importance sampling distributions propose histories which contribute most to the sum. But these distributions 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 which 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.

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 :

$$N(t) = \begin{cases} N_0 \left( \frac{2m}{N_{anc}} \right)^D & \text{si } 0 \leq t \leq D \\ N_{anc} & \text{si } t \geq D. \end{cases}$$

Likelihood of the data

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

$$\text{Prob}(\{n_0, \ldots, n_s\} | \H) = \int_{\H} \prod_{j=1}^{m} p_{\theta} (n_{anc}) \left[ \prod_{i=1}^{s} f(s_i | n_{anc}) \right] d\H$$

$$= \int \g(\{n_0, \ldots, n_s\}) d\H.$$

Where :
- $n_{anc}$: observed data,
- $(n_0, \ldots, n_{s+1})$ count vector of length $(m + 2)$, such as $n_0 = n_{anc}$ and $n_{anc+1} = n_{anc} + 1$,
- $s_0, s_1, \ldots, s_s$: dates of jump (in forward time),
- $\g(\{n_0, \ldots, n_s\}) = \H(\H | p(\{n_0, \ldots, n_s\}))$,
- $\H$: 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 increase multiple copies of good histories, to generate futur better histories.

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 $\theta_{anc} = 400$.

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.

<table>
<thead>
<tr>
<th></th>
<th>SIS SISR</th>
<th></th>
<th>SIS SISR</th>
<th></th>
<th>SIS SISR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel. bias $\theta$</td>
<td>0.56</td>
<td>0.364</td>
<td>4.92</td>
<td>1.62</td>
<td>1.35</td>
</tr>
<tr>
<td>$D$</td>
<td>-0.0201</td>
<td>0.0308</td>
<td>-0.0608</td>
<td>0.177</td>
<td>0.191</td>
</tr>
<tr>
<td>$\theta_{anc}$</td>
<td>0.0479</td>
<td>-0.138</td>
<td>0.0438</td>
<td>-0.00967</td>
<td>0.6522</td>
</tr>
<tr>
<td>RMSE $\theta$</td>
<td>0.711</td>
<td>0.557</td>
<td>5.17</td>
<td>1.82</td>
<td>2.98</td>
</tr>
<tr>
<td>$D$</td>
<td>0.142</td>
<td>0.142</td>
<td>0.141</td>
<td>0.417</td>
<td>0.442</td>
</tr>
<tr>
<td>$\theta_{anc}$</td>
<td>0.369</td>
<td>0.305</td>
<td>0.245</td>
<td>0.21</td>
<td>0.322</td>
</tr>
<tr>
<td>With $\theta = 0.4$, $D = 1.25$ and $\theta_{anc} = 400$.</td>
<td>With $\theta = 0.4$, $D = 0.25$ and $\theta_{anc} = 400$.</td>
<td>With $\theta = 0.4$, $D = 0.25$ and $\theta_{anc} = 400$.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Acknowledgments

References