Ré-échantillonnage dans un shéma séquentiel d'échantillonnage préférentiel

Coralie Merle [1,2,3]

[1] *Institut de Mathématiques et de Modélisation de Montpellier (I3M), Université de Montpellier, coralie.merle@univ-montp2.fr*
[2] *Centre de Biologie pour la Gestion des Populations (CBGP), INRA*
[3] *Institut de Biologie Computationnelle (IBC), Montpellier*

Résumé. Nous nous intéressons au calcul de la vraisemblance d'un modèle à processus latent pour une valeur ϕ fixée du paramètre d'intérêt. Nous appliquons une méthode d'échantillonnage préférentiel sur les trajectoires d'un processus Markovien de saut inhomogène en temps jusqu'à un temps d'arrêt τ. Pour améliorer cet échantillonnage de l'espace des trajectoires, avant d'atteindre le temps d'arrêt, nous proposons de ré-échantillonner les débuts des trajectoires en fonction des poids et de l'état courant. Nous expliquerons quand et comment ré-échantillonner. Les méthodes d'échantillonnage préférentiel sont particulièrement utilisées en génétique des populations. En effet, la distribution du polymorphisme génétique d'un échantillon actuel dépend de l'évolution de la taille de la population au travers de processus stochastiques latents : son histoire passée. Mais ces méthodes sans ré-échantillonnage ne sont pas toujours efficaces, en particulier pour des modèles de populations dont la taille varie au cours du temps. Nous mettrons en évidence le gain obtenu grâce au ré-échantillonnage sur le cas d'une contraction de la taille de la population.

Mots-clés. Échantillonnage préférentiel, ré-échantillonnage, processus Markovien de saut, génétiques des populations, inférence démographique, coalescent.

Abstract. We consider the likelihood computation of a model with a latent process for a fixed value ϕ of the parameter of interest. We apply an importance sampling method on the trajectories of a pure jump measure-valued Markov process until a stopping time τ. To improve this sampling of the trajectories space, before reaching the stopping time, we propose to resample the trajectories beginning according to the current weights and states. We describe when and how to resample. The importance sampling methods are particularly used in population genetics. Indeed the genetic polymorphism distribution of a current sample depend on the evolution of the population size through a latent process: its past history. However these methods without resampling are not always efficient, particularly for population with varying size in time. We bring out the gain obtained with the resampling technic in the case of a contracting population.

Keywords. Importance Sampling, resampling, jump Markov process, population genetics, demographic inference, coalescent.

1 Introduction

Le but de ce travail est de calculer la vraisemblance d'un modèle à processus latent pour une valeur ϕ fixée du paramètre d'intérêt. La vraisemblance des données en ϕ s'écrit alors comme l'intégrale des probabilités de chacune des trajectoires du processus compatibles avec les données observées.

On s'intéresse à une classe de méthodes de Monte-Carlo, basées sur l'échantillonnage préférentiel séquentiel (Sequential Importance Sampling : SIS). Dans ce shéma, la distribution d'importance propose des trajectoires du processus parmi celles qui contribuent le plus à la somme définissant la vraisemblance.

Ces méthodes sont utilisées en génétiques des populations, voir notamment (Griffiths et Tavare,1994, Stephens et Donnelly, 2000, De Iorio et Griffiths, 2004a et 2004b, De Iorio *et al.*, 2005). En effet, la distribution du polymorphisme génétique d'un échantillon d'individus dépend de l'évolution de la taille de la population au travers de processus stochastiques (modèles de Wright-Fisher, généalogie de Kingman, etc.) non observés. Dans ce cadre, le processus latent est l'histoire (généalogie avec mutation) de l'échantillon observé. Mais les méthodes d'échantillonnage préférentiel ne sont pas toujours efficaces, en particulier pour des modèles de populations en déséquilibre (Leblois *et al.*, 2014) et le temps de calcul augmente fortement pour la même précision de l'estimation de la vraisemblance en un point de l'espace des paramètres.

Nous avons exploré l'échantillonnage préférentiel séquentiel avec ré-echantillonnage (Sequential Important Sampling and Resampling : SISR). L'idée est de ré-chantillonner, au cours de la construction des trajectoires, de façon à apprendre progressivement quelles sont les trajectoires proposées par la distribution d'échantillonnage qui contribuent le plus à la somme. On espère ainsi diminuer le temps de calcul. Nous répondrons aux questions suivantes :

- Quand ré-chantillonner ? (À quels moments dans les trajectoires ?)
- Comment ré-chantillonner ? (Suivant quels poids ?)

Nous appliquerons cette nouvelle procédure en génétique des populations et mettrons en évidence le gain obtenu grâce au ré-échantillonnage sur le cas d'une vraisemblance modélisant une contraction de la taille de la population.

2 Échantillonnage préférentiel séquentiel

Calcul de la vraisemblance par échantillonnage préférentiel On considère un modèle décrit par un processus markovien de saut pur $Z(t)$, à valeur dans l'espace des mesures de comptage. À chaque date t, $Z(t)$ donne la distribution allélique des ancêtres

de l'échantillon. On s'intéresse plus particulièrement au cas où ce processus markovien est inhomogène en temps : les taux de transition dépendent de l'état de la trajectoire au temps t et de t.
Ainsi, si $X = \{X_k\}_{k\in\mathbb{N}}$ est la chaîne de Markov incluse du processus markovien de saut, si $\tau + 1$ est un temps d'arrêt et si $\mathbf{n}_{\text{obs}}$ est la mesure de comptage observée à l'instant τ, c'est à dire les données observées, la vraisemblance du paramètre ϕ est

$$L(\mathbf{n}_{\text{obs}}|\phi) = \mathbb{P}_\phi\left(\mathbf{X}_\tau = \mathbf{n}_{\text{obs}}\right). \tag{1}$$

Elle est obtenue en sommant sur toutes les trajectoires possibles (latentes) :

$$L(\mathbf{n}_{\text{obs}}|\phi) = \sum_{k\geq 1} \sum_{H\in\mathcal{H}^k_{\mathbf{n}_{\text{obs}}}} \mathbb{P}_\phi\left(H\right), \tag{2}$$

où $\mathcal{H}^k_{\mathbf{n}_{\text{obs}}}$ est l'espace des trajectoires de longueur k (i.e. $\tau = k$) compatibles avec les données observées ($H_k = \mathbf{n}_{\text{obs}}$) et P la distribution de notre processus.

L'idée de l'échantillonnage préférentiel est de changer la distribution d'échantillonnage sur l'espace des trajectoires $\mathcal{H}$. Si on note $\mathrm{P} = (p(x,y))$ la matrice transition de la chaîne et si $\mathrm{Q} = (q(x,y))$ est une autre matrice de transition sur $\mathcal{H}$ telle que $p(x,y) = 0$ dès que $q(x,y) = 0$ et $Q(H) = 0$ si $H \notin \cup_k \mathcal{H}^k_{\mathbf{n}_{\text{obs}}}$, alors on obtient

$$L(\mathbf{n}_{\text{obs}}|\phi) = \mathbb{E}_{H\sim\mathrm{Q}}\left[P\left(H^{(j)}\right)/\mathrm{Q}\left(H^{(j)}\right)\right]. \tag{3}$$

La vraisemblance est alors une espérance sous la distribution Q, on peut donc l'estimer par un estimateur de type Monte-Carlo :

$$\widehat{L(\mathbf{n}_{\text{obs}}|\phi)} = \frac{1}{n_H}\sum_{j=1}^{n_H} w^{(j)}, \quad \text{où} \quad w^{(j)} = \frac{P\left(H^{(j)}\right)}{\mathrm{Q}\left(H^{(j)}\right)} = \prod_{i=1}^{k} \frac{p(H^{(j)}_{i-1}, H^{(j)}_i)}{q(H^{(j)}_{i-1}, H^{(j)}_i)} \tag{4}$$

est le poids d'importance de la j-ième trajectoire $H^{(j)}$ tirée selon la loi d'importance Q.

On rappelle que l'échantillonnage préférentiel peut devenir inefficace dans des espaces de grande dimension car la variance du rapport de vraisemblance tend vers l'infini exponentiellement vite avec la longueur de la trajectoire (voir, par exemple, Asmussen et Glynn (2007) V.1.17 et XIV.5.5). C'est un problème bien connu de l'échantillonnage préférentiel séquentiel pour de longues réalisations de la chaîne de Markov.

3 Ré-échantillonnage

Motivation Dans des modèles simples, on dispose de distributions d'importance efficaces (voir notamment De Iorio *et al.*, 2005). L'inhomogénéité en temps du processus considéré ici rend ces distributions d'importance Q inefficaces. En appliquant une technique de ré-échantillonnage, nous visons à améliorer l'efficacité de la distribution de proposition.

Le but est d'élaguer les trajectoires associées à des faibles poids d'importance et de réutiliser les débuts des trajectoires associées à des poids d'importance plus élevés.

Procédé général On arrête l'algorithme SIS (Sequential Importance Sampling) qui construit les trajectoires en parallèle à un temps donné et on modifie la composition de la collection de trajectoires selon les poids d'importance partiels à cette date. Ce nouvel algorithme est appelé SISR, voir Liu *et al.*(2001).

Supposons que, au temps k discret, on a un échantillon de n_H trajectoires pondérées par les poids partiels d'importance $S_k = \left\{(H_k^{(j)}, w_k^{(j)})\right\}_{j=1}^{n_H}$.

On implémente la stratégie de ré-échantillonnage dans laquelle on tire parmi les trajectoires courantes suivant une distribution multinomiale où l'on a mis le poids $a^{(j)}$ à la j-ième trajectoire. Ce qui donne l'algorithme suivant :

1. Pour $j' = 1, ..., n_H$,
   - $\tilde{H}_k^{(j')}$ vaut $H_k^{(j)}$ indépendamment avec probabilité proportionnelle $a^{(j)}$;
   - le poids associé à cette trajectoire est $\tilde{w}_k^{(j')} = w_k^{(j)}/a^{(j)}$.
2. Retourner la nouvelle représentation $\tilde{S}_k = \left\{(\tilde{H}_k^{(j')}, \tilde{w}_k^{(j')})\right\}_{j'=1}^{n_H}$.

Quand ré-échantillonner ? La chaîne de Markov est à valeur dans l'espace des mesures de comptage sur l'ensemble des allèles possibles.

Pour calculer notre estimateur Monte-Carlo (4) de la vraisemblance, on construit n_H trajectoires en parallèle, partant toutes du même état initial $\mathbf{n}_{\text{obs}}$.

Nous cherchons à quel moment de la construction des trajectoires nous allons ré-échantillonner. Le but est de pouvoir discriminer les trajectoires qui seront les moins probables à la fin de leur construction.

Une première idée serait de choisir le moment où nous allons ré-échantillonner après un nombre donné de transitions.

Les états courants des trajectoires ne sont pas de même nature (les mesures de comptage n'ont pas la même masse totale) et ne sont pas facilement comparables.
Un choix plus efficace consiste à ré-échantillonner les trajectoires aux moments où les mesures de comptage représentant les états courants ont même masse totale. Alors, les états courants des trajectoires sont facilement comparables.

Comment choisir les poids de ré-échantillonnage ? Notre idée est de choisir les $a^{(j)}$ de façon à refléter une tendance future des trajectoires, plus précisément de façon à favoriser les trajectoires qui auront un poids d'importance final élevé. Évidemment on ne

connait pas le poids final de chaque trajectoire mais on peut l'approcher même vaguement par une fonction qui dépend de l'état courant de la trajectoire, notée $f(H_k^{(j)}, \beta)$, et en tenant compte du poids courant $w_k^{(j)}$ de la trajectoire. On ré-échantillonne alors suivant les probabilités :

$$a^{(j)} = \left[w_k^{(j)}\right]^{\alpha} f(H_k, \beta),$$

où $\alpha \in [0, 1]$ et $\beta \geq 0$ sont des paramètres de calibrage.

4 Applications en génétique des populations

Nous présenterons des résultats numériques sur un modèle démographique de contraction de la taille de la population au cours du temps.

La généalogie (i.e., les liens de parenté de l'échantillon de copies de gènes étudié), les dates des mutations et les génotypes ancestraux, décrits par le modèle stochastique, ne sont pas observés directement. Ils représentent le processus latent. La vraisemblance des données s'obtient alors en sommant sur toutes les possibilités de ce processus latent.

Nous comparerons les deux procédures d'inférence (SIS et SISR) de la vraisemblance sur des jeux de données simulés et différentes calibrations possibles de α et β.

En particulier nous montrerons une diminution de l'erreur quadratique moyenne pour l'estimation de la vraisemblance lorque l'on utilise la procédure avec ré-échantillonnage.

Remerciements

Je remercie Jean-Michel Marin, Raphaël Leblois, Pierre Pudlo et François Rousset pour leur encadrement dans ce travail.
L'auteur a été financièrement soutenu par le LabEx NUMEV (Solutions Numériques, Matérielles et Modélisation pour l'Environnement et le Vivant, ANR-10-LABX-20) et le LabEx CeMEB (Centre Méditerranéen de l'Environnement et de la Biodiversité).

Bibliographie

[1] Søren Asmussen et Peter W Glynn (2007), *Stochastic simulation: Algorithms and analysis*, volume 57, Springer Science & Business Media.
[2] Robert C Griffiths et Simon Tavare (1994), *Sampling theory for neutral alleles in a varying environment*, Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 344(1310):403–410.
[3] Maria De Iorio et Robert C Griffiths (2004a), *Importance sampling on coalescent histories. i*, Advances in Applied Probability, pages 417–433.

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