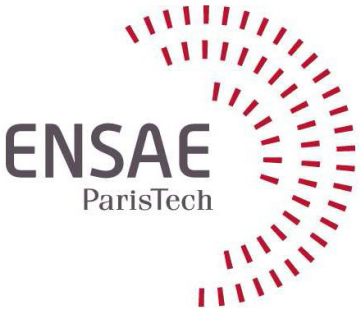


INSTITUT POLYTECHNIQUE DE PARIS

Hidden Markov Models and Sequential Monte Carlo

A rare event approach to high dimensional Approximate Bayesian Computation

Basé sur un article de : Dennis Prangle, Richard G. Everitt et Theodore Kypraios
Auteurs : Maxime Berillon, Tristan Legris, Alexandre Marquis
Professeur: Nicolas Chopin

January 6, 2021

Contents

1 Introduction

Le but de ce projet est d'étudier l'article *A rare event approach to high dimensional Approximate Bayesian computation* de Dennis Prangle, Richard G. Everitt et Theodore Kypraios. Nous avons dû analyser et comprendre les enjeux débouchant à la création des algorithmes que nous allons vous présenter ainsi que leur fonctionnement, dans le but de vous en expliquer leur principe et de vous présenter une application possible (dans le milieu épidémiologique) ainsi que des résultats pratiques que nous avons pu en tirer. La section 2 propose une brève description du modèle proposé. Dans la section 3 nous montrons comment nous avons implémenté le modèle proposé, dans un exemple gaussien pui avec des données réelles. Dans la section 4 nous concluons avec une discussion et une conclusion.

2 Modèle

Le modèle proposé par Prangle et al. comprend plusieurs algorithmes qui travaillent ensemble.

2.1 Pseudo-Marginal Metropolis-Hastings — RE-ABC

L'algorithme Pseudo-Marginal Metropolis-Hastings (PMMH) est une méthode de Monte Carlo dérivée de l'algorithme de Metropolis-Hastings. Cet algorithme (MH) est utilisé pour tirer une séquence d'observations selon une loi P lorsqu'il n'est pas possible de simuler directement ; il faut pour cela pouvoir disposer d'une fonction f proportionnelle à la densité cible de P. La différence du PMMH est de pouvoir continuer de simuler selon cette même loi en remplaçant f par une estimation $\widehat{f}$. Dans le cas qui nous intéresse on remplace la vraisemblance ABC par son estimation.

Ainsi à chaque étape t on propose θ' simulé selon une loi normale centrée en θ_{t-1} et u simulé selon une loi uniforme $\mathcal{U}(0,1)$. On calcule

$$p = \frac{\pi(\theta')\widehat{L}'_{ABC}q(\theta_{t-1}|\theta')}{\pi(\theta_{t-1})\widehat{L}_{ABC,t-1}q(\theta'|\theta_{t-1})}$$

Si $u > p$, θ' est rejeté, sinon il est accepté. Ainsi en sortie le RE-ABC renvoie une séquence de $\theta_1, ..., \theta_T$ simulés selon la loi a posteriori $\pi_{ABC}(\theta|y)$.

2.2 Rare Event Sequential Monte Carlo — RE-SMC

Nous venons de le voir, le RE-ABC a besoin d'une estimation de $\widehat{L}_{ABC}$, l'algorithme RE-SMC permet de fournir cette estimation. Les algorithmes Rare Event Sequential Monte Carlo ont été proposé par Cérou et al. en 2012 [1]. Le but principal de ces filtres particulaire est d'estimer de toutes petites probabilités. Il permettent entre autre l'estimation de la vraisemblance de manière plus précise que les méthodes ABC habituelles (ABC rejection sampling par exemple). On distingue deux algorithmes FIXED-RE-SMC (dans lequel la suite des seuils $\epsilon_1, ...,_T$ sont fixés) et ADAPT-RE-SMC (dans lequel les seuils $\epsilon_1, ...,_T$ sont calculés par l'algorithme à chaque étape, seul un seuil cible est renseigné).

A chaque étape de ce filtre particulaire on applique un kernel markovien de densité invariante $\pi(x|\theta, \Phi(x) \leq \epsilon_{t-1}$. Ce kernel est décrit dans la section suivante.

2.3 Slice Sampling

Le kernel utilisé est le *Slice Sampling*. À chaque étape du RE-SMC on applique cette transformation aux particules survivantes. Très schematiquement on applique une perturbation à la particule retenue x en vérifiant que le résultat x' vérifier toujours $\Phi(x') \leq \epsilon$.

Nous verrons plus loin quelle fonction Φ est utilisée.

3 Applications

3.1 Implémentation de la solution

L'algorithme de simulation a été implémenté à l'aide du package `particles`. Deux classes `fixed-RE-FK` et `adapt-RE-FK` héritée de la classe mère permettent d'introduire les spécificités du sujet qui nous occupe. Le noyau repose sur le slice sampling, les poids sont nuls ou égaux à 1 et le nombre d'étape dépend de la séquence (ϵ_t) choisie. L'héritage permet d'exploiter les propriétés de la classe SMC. Enfin, l'algorithme ABC repose sur la classe du même nom.

3.2 Application gaussienne

Dans un premier temps on applique la théorie à des données simulées suivant la même démarche que l'article. Un vecteur Y est tiré suivant une loi gaussienne centrée de variance $\sigma = 3$. Les particules représentent les variables latentes X qui sont donc les quantiles des données simulées à chaque étape du SMC. La figure 1 illustre le bon fonctionnement de l'algorithme. Dans cet exemple on teste un sigma proposé égal à 5. La variance étant trop élevée on s'attend à ce que les variables latentes (et donc leurs images) se resserrent afin d'approcher les données Y, ce que l'on constate effectivement (ici sur la $10^{\text{ème}}$ coordonnée de chaque particule, avec les marges à 5% et 95%).

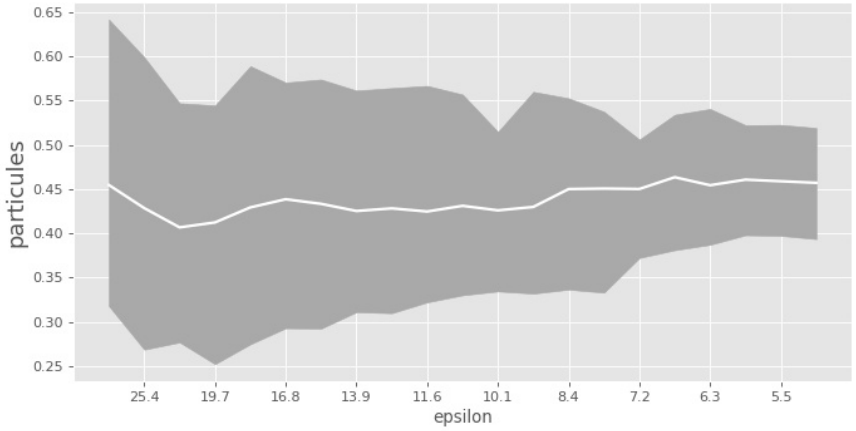


Figure 1: Convergence des particules dans RE SMC

La séquence (ϵ_t) est choisie d'après le conseil des chercheurs : on lance d'abord la méthode ADAPT-RE-SMC afin d'obtenir un nombre d'étape adaptée. En effet si la séquence choisie est trop longue alors on risque de perdre du temps inutilement, si elle est trop courte on court le risque que l'ESS s'annule brutalement. A ce titre, l'article suggère de fixer cette séquence pour un paramètre proche du paramètre estimé a priori ce qui est étonnant si l'on considère que l'évaluation des cas de faible probabilité nécessite un RE-SMC plus progressif. On génère donc les (ϵ_t) pour $\sigma_{proposé} = 8$.

Outre ϵ, il est également très important d'ajuster le nombre de particules nécessaires. D'après les informations disponibles, les simulations de l'article ont été menées avec quelques dizaines de particules, entre 20 et 50. Il est suggéré de choisir ce nombre en fonction de la variance souhaitée. La figure 2 présente l'évolution de la variance fonction du nombre de particules et du "threshold", c'est à dire le dernier epsilon de la suite.

On observe clairement que la variance atteint un plateau à partir de 50 particules qui est donc le nombre que l'on retiendra dans la suite des simulations. De plus lorsque le threshold est faible la variance l'est également. Cependant, on cherche à ce qu'epsilon soit aussi petit que possible dans la limite du temps de calcul disponible. On gardera donc un seuil cible de 5 dans RE ABC, ce qui permet notamment de gagner en précision dans les zones de fortes probabilités du PMMH.

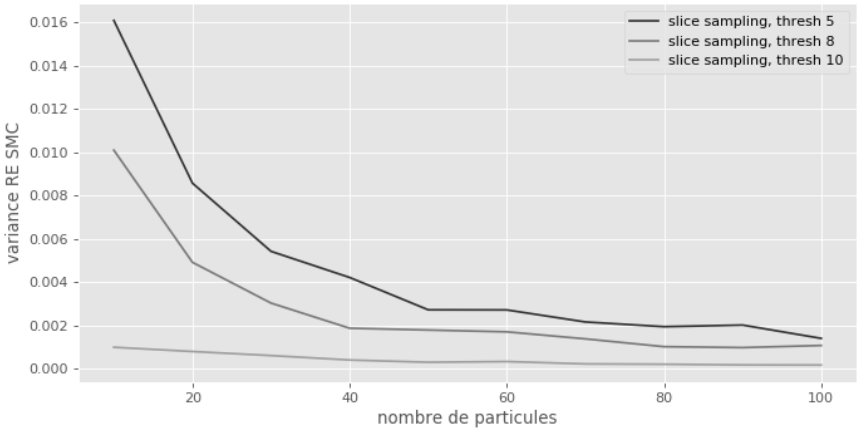
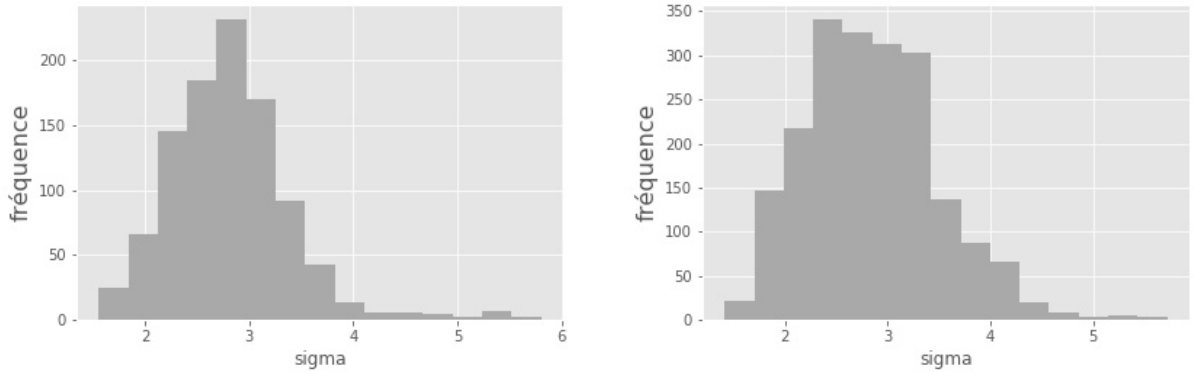


Figure 2: Variances RE SMC en fonction des paramètres

On s'intéresse maintenant à l'algorithme RE ABC qui implémente un PMMH pour le cas des évènements rares. Le papier suggère de l'initialiser en une valeur de haute vraissemblance pour éviter une période "burn-in". Le prior est une uniforme sur $[0, 10]$. On présente sur la figure 3 la distribution a posteriori pour deux variances de la distribution "proposal".

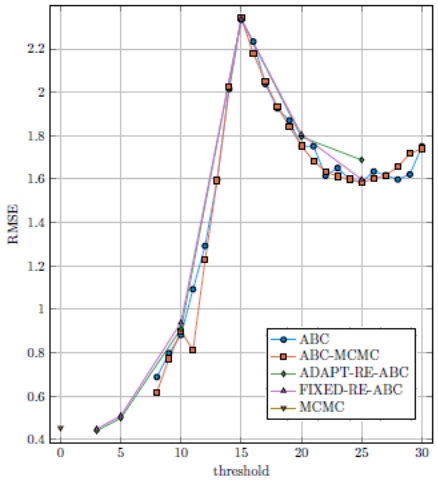


(a) Variance de la méthode de proposition simple (b) Variance de la méthode de proposition paramétrisée

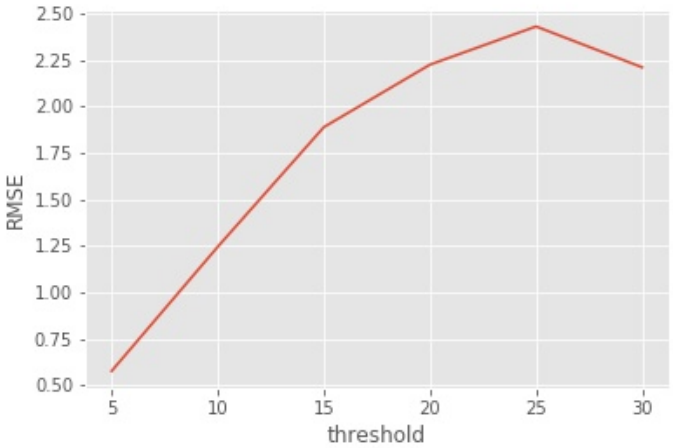
Figure 3: Distribution a posteriori RE ABC

Dans le premier cas la variance a été choisie empiriquement afin de trouver la variance a posteriori de la distribution. La littérature sur PMMH propose de fixer la variance de q $\frac{2.562^2}{dim(\theta)}\hat{\Sigma}$. C'est ce qui est fait sur le graphique 3b.La variance est fixée à $2.562^2 \times 0.34 = 2.25$. Cependant le résultat n'est pas vraiment probant. Les variances a posteriori ne sont pas similaires, de plus la variance plus élevée génère un taux de rejet (63%) supérieur au premier cas (43%) ce qui impose de simuler plus de fois (2000 simulations à droite et 1000 à gauche).

Afin d'évaluer la performance de notre ABC on évalue la RMSE de la distribution a posteriori pour le cas FIXED RE SMC et on compare avec les résultats de l'article (figure 4). On retrouve des erreurs comparables mais la forme de la courbe ne correspond pas. Il serait nécessaire de prolonger cette étude pour plus avant pour retrouver les résultats.



(a) RMSE de l'article



(b) RMSE obtenue pour fixed abc

Figure 4: RMSE de PMMH

3.3 Application épidémiologique

3.3.1 Modélisation

SIR C'est un modèle très utilisé permettant de modéliser une maladie infectieuse en représentant trois groupes distincts : les ***S****usceptibles* (individus sains pouvant être contaminés par la maladie), ***I****nfectious* (individus infectés pourvant transmettre la maladie) et enfin ***R****emoved* (individus ultérieurement infectés mais ne pouvant plus l'être car immunisées, mortes ou en quarantaine).

Dans ce modèle on s'intéresse à deux paramètres : λ le taux d'infection (plus il est élevé plus la maladie est contagieuse) et γ le taux de suppression (plus il est élevé plus les temps d'infections sont courts). Enfin on s'intéresse à leur rapport $R_0 = \frac{\lambda}{\gamma}$, le nombre de reproduction de base. Si ce dernier est inférieur à 1 l'agent pathogène infectera moins d'une personne en moyenne par cas, et finira par disparaître.

Construction de Sellke L'algorithme de construction de Sellke (Sellke, 1983) permet de simuler une épidémie à partir de deux vecteurs de variables $g_1, ..., g_n$ et $p_1, ..., p_n$.

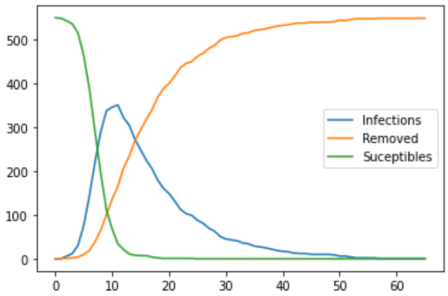
g_i est la période d'infection, c'est à dire la durée durant laquelle l'individu peut transmettre la maladie. Dans notre cas, les g_i suivent une loi exponentielle de paramètre λ (le taux d'infection de la maladie).

p_i est le seuil de pression : la pression exercée par la maladie sur un individu augmente à mesure qu'augmente le nombre d'individus infectés. Formmellement elle est définie comme $p = \frac{\lambda}{n_{total}} \times n_{infected}$. Lorsque cette pression est supérieure à un seuil p_i, l'individu i est infecté. Dans notre cas les g_i sont tirés indépendamment selon une loi exponentielle de paramètre 1.

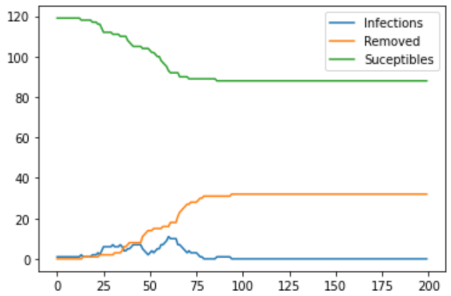
3.3.2 Données Abakaliki

Le cas réel présenté dans l'article de Prangle et al. se base sur un jeu de données recueilli par Thompson et Foege en 1967 lorsqu'ils étudiaient l'évolution de la propagation de la variole au sein de la ville d'Abakaliki (situé au Nigeria) [3]. Ce jeu de données regroupe des informations concernant 32 personnes infectées au sein d'une population isolée d'une taille totale de 120 individus.

Le dataset comporte la date à laquelle sont apparus les symptômes mais pas la date à laquelle les individus ont été *removed*. Afin de modéliser la période d'infection, nous avons réutilisé les hypothèse formulées par Jessica Stockdale et al. ainsi que leur résultats numériques (voir Jessica Stockdal et al., 2017[2]). Nous avons donc supposé que la distribution des périodes infectieuses suivaient une loi gamma dont la moyenne



(a) Simulation d'un modèle SIR grâce à l'algorithme de construction de Sellke



(b) Modèle SIR du dataset Abakaliki

et la déviation standard ont été calculées en fonction des cas observés, et ignoré la possibilité de mise en quarantaine des individus puisque nous n'avions pas assez de données concernant ce paramètre.

Nous obtenons ainsi $r_{(1)} < ... < r_{(n)}$ les *removal times* des n individus de la population que l'on transforme en $s_{(i)}$ qui représentent le temps depuis le premier *removal*.

3.3.3 Implémentation

Tout l'enjeu ici est d'obtenir une approximation du R_0 de l'épidémie qui a touché la ville d'Abakaliki.

Paramétrisation Nous utilisons l'algorithme RE-ABC proposé par Prangle et al. avec comme paramétrisation $\theta = (\lambda, \gamma)$. Tous deux se voient attribués une *prior* peu informative $Exp(10)$ afin d'orienter le RE-ABC vers des valeurs faibles de ces deux paramètres.

Particules Les particules utilisées par le RE-SMC seront les quantiles des variables $g_1, ..., g_n$, $p_1, ..., p_n$ ce qui permet d'utiliser le *Slice Sampling* (nécessité d'être distribuées uniformément a priori). Ces particules seront les variables latentes qui permettront, avec θ, de simuler une épidémie.

Fonction Φ La fonction Φ, utilisée dans le *Slice Sampling*, permet de mesurer la distance entre la simulation issue de la particule x et le dataset de départ. On a $\Phi(x) = d(y(x), y)$. d est la fonction de distance définie par Prangle et al. : distance de Minkowski avec $p = 2$ et une pénalisation pour tout *mismatch* dans le nombre d'infectés. $y(x)$ représente le dataset simulé, via l'algorithme de construction de Sellke, à l'aide de la particule x et en utilisant le θ courant.

3.3.4 Résultats

Nous présenterons ici des résultats obtenus avec un dataset simulé grâce à l'algorithme de construction de Sellke. Le jeu de données Abakaliki est très large et les temps de calcul très long pour produire un nombre suffisant d'itérations avec le RE-ABC. En réduisant la population à 5 sur le dataset simulé nous avons plus de marge de manoeuvre pour paramétriser le modèle.

Nous générons un jeu de données avec $\lambda = 0.075$ et $\gamma = 0.0455$ soit un $R_0 = 1.65$. Avec 300 itérations du RE-ABC nous obtenons les résultats suivant :

Modèle	Estimation de λ	Écart-type de λ	Estimation de γ	Écart-type de γ	Estimation de R_0
SIR	0.071	0.11	0.044	0.10	1.61

Figure 6: Résultat de l'estimation du R_0 sur un jeu de données simulées

4 Discussion et conclusion

L'élaboration de ce projet nous a permis de vraiment consolider nos connaissance sur les modèles de Markov et les méthodes de Monte Carlo séquentielles, c'est en effet en pratiquant que l'on comprend le mieux de quoi

il en retourne. Nous avons améliorer notre niveau de compréhension des modèles épidémiologiques basiques (sujet très important en ce moment). Nous avons pu mieux comprendre comment étaient programmés en pratique ce genre d'algorithmes, mais aussi quels étaient les étapes essentielles pour les optimiser en temps et en ressource. En faisant tourner ces derniers de nombreuses fois avec différents paramètres, nous avons mieux cerné comment évoluaient leurs résultats en fonction des paramètres d'entrée. Nous avons cependant dû beaucoup nous reposer sur le travail des auteurs afin d'obtenir de bons résultats.
Il est néanmoins possible d'aller plus loin et de creuser un peu plus le sujet, nous n'avons par exemple pas exploré la possibilité d'utiliser un loi de Weibull pour les seuil de pression du modèle SIR, ...

References

[1] Frédéric Cérou, Pierre Del Moral, Teddy Furon, and Arnaud Guyader. Sequential monte carlo for rare event estimation. *Statistics and computing*, 22(3):795–808, 2012.

[2] Jessica E. Stockdale, Theodore Kypraios, and Philip D. O'Neill. Modelling and bayesian analysis of the abakaliki smallpox data. *Epidemics*, 19:13 – 23, 2017.

[3] David Thompson, W. H Foege, and National Communicable Disease Center (U.S.). Faith tabernacle smallpox epidemic, abakaliki, nigeria / by david thompson and william foege, 1968.