Méthodes ABC en épidémiologie

Clément Dell'Aiera

Résumé

Nous présentons dans ce court rapport une introduction aux méthodes ABC, avec application au modèle SIR incorporant le contact tracing. Nous suivons l'articlede Blum et Tran [2] pour l'application en épidémiologie, ainsi que l'article de Beaumont [1]pour la présentation des méthodes ABC.

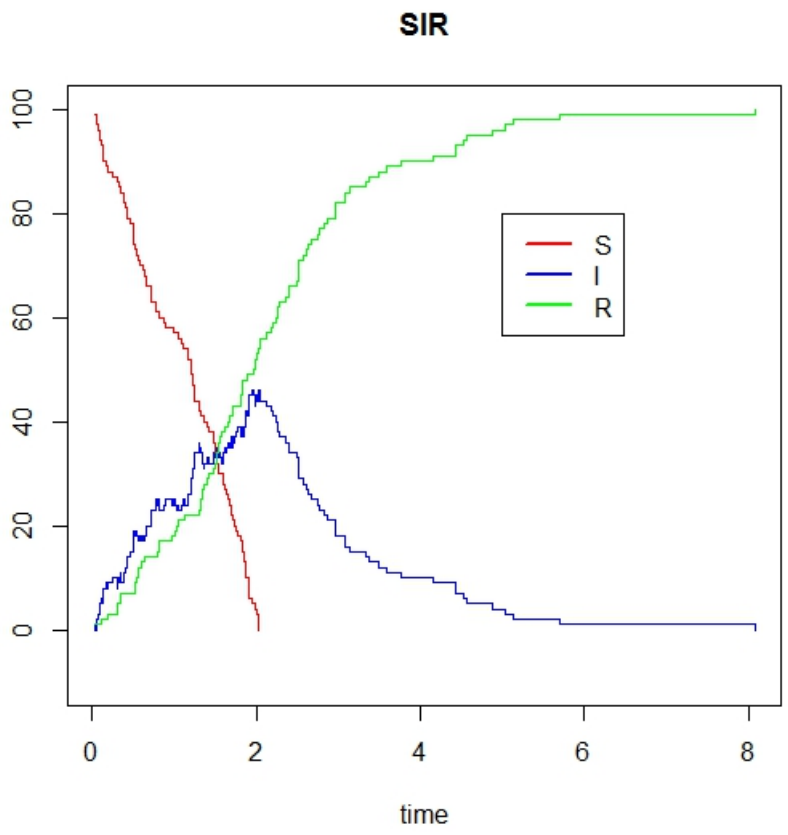

FIGURE 1 – Simulation du modèle SIR sur une population de 100 individus.

Table des matières

1 Principe et historique

1.1 Méthodes ABC.

Rappelons le principe de l'inférence bayésienne : le modélisateur se donne un prior π sur le paramètre θ et une famille de loi $p(x|\theta)$. Suite à des observations x, on met à jour la loi, i.e. on cherche à calculer :

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)})$$

où : $p(x) = \int p(x|\theta)\pi(\theta)d\theta$. Cette dernière intégrale peut être difficile à calculer numériquement, typiquement si l'espace des paramètre Θ est de grande dimension, mais cette difficulté peut être traitée avec des algorithmes classiques tels que Monte Carlo Markov Chains par exemple, où les modélisateur a seulement besoin de connaître la loi a posteriori à une constante de normalisation près, pour pouvoir simuler des paramètres selon des données observées.

Ces méthodes que nous avons appelées "classiques" comportent néanmoins un inconvénient majeur : l'expérimentateur n'a aucun contrôle sur l'erreur commise dans l'approximation de sa loi p par des X_n , i.e. l'entier :

$$N_\epsilon = \inf_n \{\forall m \geq n, |P(X_m = \theta) - p(x)| < \epsilon\}$$

est bien défini, mais n'est pas calculable de façon universelle.

Toutefois si la loi est discrète et de basse dimension, Rubin [1] a proposé un algorithme qui permet de simuler des paramètres à partir de données sans vraisemblance.

Soit y une observation, tant que le nombre de simulations acceptées est inférieur à N, répéter :

1. Simuler $\theta_i \sim \pi(\theta)$
2. Simuler $x_j \sim p(x|\theta)$
3. Si $x_j \neq y$, rejeter x_j.

On voit que si les données suivent une loi continue, la probabilité que l'algorithme accepte un nombre non nul de simulation est 0. Rubin a alors proposé une modification de l'étape 3 basé sur une discrépance ρ qui fait office de distance entre les observations et les points simulés. Pour illustrer par un exemple simple, on peut imaginer que ρ est la distance euclidienne usuelle.

Soit y une observation, tant que le nombre de simulations acceptées est inférieur à N, répéter :

1. Simuler $\theta_i \sim \pi(\theta)$
2. Simuler $x_j \sim p(x|\theta)$
3. Si $\rho(x_j, y) > \epsilon$, rejeter x_j.

Pour s'attaquer aux grandes dimensions, on peut se servir d'une application S à valeur dans un espace de petite dimension et aboutir à une autre version de ce même algorithme. On impoose à S de vérifier : $p(\theta|x) = p(\theta|S(x)), \forall \pi$ afin de limiter la perte d'information : cette application S peut être pensée comme une statistique exhaustive pour le modèle.

Soit y une observation, tant que le nombre de simulations acceptées est inférieur à N, répéter :

1. Simuler $\theta_i \sim \pi(\theta)$
2. Simuler $x_j \sim p(x|\theta)$
3. Si $\rho(S(x_j), S(y)) > \epsilon$, rejeter x_j.

Pour résumer, la méthode ABC (Approximate Bayesian Computing) propose, si l'on veut estimer la distribution *a posteriori* d'un paramètre θ, de simuler des variables θ_j selon notre *prior* π, à partir desquelles sont simulées des échantillons y_j, j allant de 1 à n le nombre de simulations. Un ensemble de *summary* $S(y_j)$ est alors calculé à partir des données simulées, quantités que l'on compare à la valeur $S(y_0)$ de S sur le véritable échantillon observé. Si la "distance" entre $S(y_j)$ et $S(y_0)$ ne dépasse pas un certain seuil ϵ fixé à l'avance, on garde la simulation θ_j, sinon on la rejette.

1.2 Un exemple simple

Nous allons illustrer le dernier algorithme sur un exemple simple. On se place dans le cas d'un modèle bayésien avec un prior normal $\pi(\theta) \sim \mathcal{N}(0, 1)$ et d'une loi $p(x|\theta) \sim \mathcal{N}(\theta, 1)$. On sait que les lois normales sont conjuguées :

$$p(\theta|x) \sim \mathcal{N}(x, \frac{1}{2}).$$

Regardons ce que Scilab nous donne à partir de l'observation $x = 1$, de 5000 simulations (acceptées) et de $\epsilon = 0.01$.

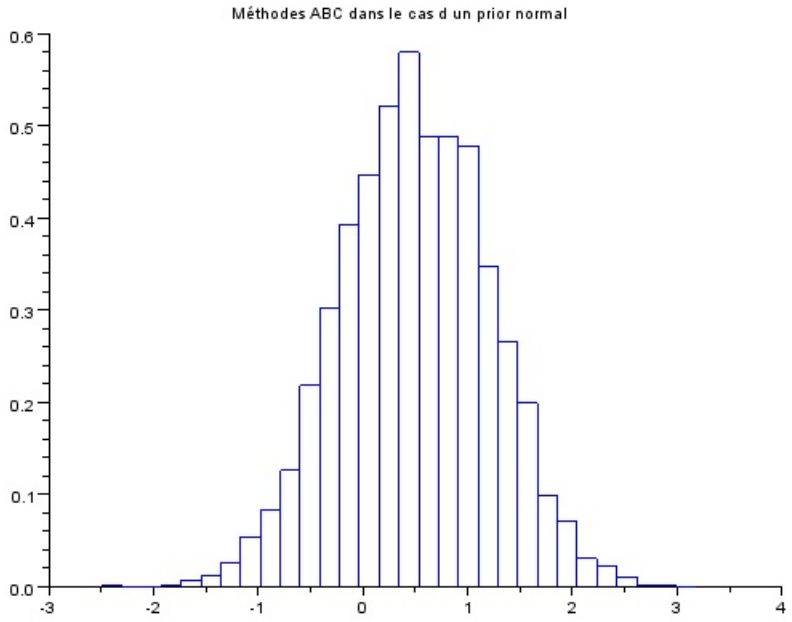

FIGURE 2 – Histogramme de 5000 simulations de θ

2 Trois algorithmes importants en méthode ABC

2.1 Correction par régression locale

Nous l'avons vu, la méthode ABC ne permet pas de simuler exactement la loi désirée. La méthode de correction par régression locale propose de corriger cet écart en utilisant le modèle de régression :

$$\theta_i = f(S(x_i)) + \epsilon_i$$

où f est la fonction de régression et les ϵ_i sont centrés de même variance. On note y la valeur observée. Le principe est le suivant : on donne un poids plus fort aux simulations qui sont le plus proches de $S(y)$. Pour cela, à chaque simulation $(\theta_i, S(x_i))$ est associé le poids :

$$K(\rho(S(x_i), S(y))$$

où l'on s'est fixé le noyau K à l'avance. Une fois la régression effecuée, on obtient un échantillon pondéré de la loi *a posteriori* en corrigeant les θ_i comme suit,

$$\theta_i^* = \hat{f}(S(y)) + \frac{\hat{\sigma}(S(y))}{\hat{\sigma}(S(x_i))}\hat{\epsilon}_i$$

où $\hat{f}$ est l'ésperance conditionnelle estimée, les $\hat{\epsilon}_i$ sont les résidus empiriques de la régression, et $\hat{\sigma}$ dénote l'écart-type empirique conditionnel.

Voici un algorithme qui se base sur cette méthode, issu de l'article de Beaumont [1].

1. Si y est une observation, tant que le nombre de simulations est inférieur à M :
   (a) Simuler $\theta_i \sim \pi(\theta)$
   (b) Simuler $x_i \sim p(x|\theta_i)$
2. Soit $k_j \leftarrow$ l'écart type empirique des $S_j(x)$
3. $\rho(S(x), S(y)) := \sqrt{\sum_{j=1}^{s}(S_j(x) - S_j(y))^2}$
4. Choisir la tolérance ϵ telle que la proportion de points acceptés P_ϵ vaille N/M
5. Pondérer les simulations $S(x_i)$ avec $K_\epsilon(\rho(S(x_i), S(y)))$, où
$$K_\epsilon(t) = \epsilon^{-1}(1 - (t/\epsilon)^2)1_{t\leq\epsilon}(t)$$
6. Estimer $\hat{E}[\theta|S(x)]$ grâce à une régression linéaire pondérée appliquée aux N points de poids non nuls
7. Ajuster $\theta_i^* = \theta_i - \hat{E}[\theta|S(x_i)] + \hat{E}[\theta|S(y)]$
8. Les θ_i^* obtenus, de poids $K_\epsilon(\rho(S(x_i), S(y)))$, sont des simulations approchées de la loi *a posteriori* $p(\theta|y)$

2.2 Algorithme MCMC dans le cadre ABC

Dans cette section, on détaille l'algorithme MCMC adapté au cadre ABC. Un inconvenient des méthodes par rejet et régression ABC (que nous illustrerons dans la partie 2) est que l'on simule les paramètres selon le *prior* sans tenir compte de l'information qu'apporte l'échantillon. Typiquement, un échantillon qui apporterait beaucoup d'information permet, en concentrant la loi *a posteriori*, de simuler des paramètres de façon plus efficace selon cette dernière. Toutefois, le calcul de la loi *a posteriori* peut être très difficile en pratique, d'où l'idée d'utiliser les méthodes de Monte Carlo Markov Chain, que nous avons rappelées précédement. Nous suivons l'article de Beaumont pour présenter une algorithme MCMC adapté aux méthodes ABC.

On commence par simuler selon le *prior* $\pi(\theta)$, ce qui nous donne notre état initial. Pour passer d'un état à un autre, on utilise une loi que l'on choisit par exemple un noyau gaussien, que l'on note K. Ce noyau nous propose un éventuel nouvel état, que l'on accepte ou pas en fonction d'une valeur seuil.

Initialisation : Simuler $\theta_0 \sim \pi(\theta)$.
Pour n allant de 1 au nombre de simulations souhaité, faire :

1. Simuler $\theta' \sim K(\theta|\theta_{n-1})$
2. Simuler $x \sim p(x|\theta')$
3. Si $\rho(S(x)S(y)) < \epsilon$,
   (a) Simuler $u \sim \mathcal{U}_{[0;1]}$
   (b) Si $u \leq \pi(\theta')/\pi(\theta_{n-1}) \times K(\theta_{n-1}|\theta')/K(\theta'|\theta_{n-1})$
   $\theta_n = \theta'$;
   (c) Sinon $\theta_n = \theta_{n-1}$
4. Sinon $\theta_n = \theta_{n-1}$

Cette méthode présente un inconvénient : le taux d'acceptation des états peut être très faible. Il est en effet proportionnel à la fréquence à laquelle les paramètres tels que $\rho(S(x), S(y)) < \epsilon$ sont simulés, ce qui diffère des méthodes MCMC classiques, où le taux d'acceptation est proportionnel au rapport des deux vraisemblances entre état proposé et état courant. Beaumont fait remarquer dans son article qu'un mauvais choix de l'état initial par exemple loi dans la queue de la distribution de $\pi(\theta)$, peut conduire à des taux très bas, ce qui ralentit fortement les méthodes.

2.3 Méthodes de Monte-Carlo Séquentielles

Les méthodes SMC (Sequential Monte Carlo) permettent d'estimer des variables cachées x_i en observant seulement des y_j, en supposant que x soit une chaîne de Markov et que les $y_j|x$ soient indépendants. Un exemple immédiat est celui du filtre de Kalman, ou plus généralement si :

$$\begin{cases} x_{n+1} = f(x_n) + \epsilon_n \\ y_n = g(x_n) + \eta_n \end{cases}$$

où les suites ϵ et η sont par exemple des bruits blancs indépendants. On suppose que l'on connaît f,g, ainsi que la distribution de ϵ et η.

Voici un algorithme SMC adapté au cadre ABC proposé dans l'article de Beaumont [1] :

On se donne une suite décroissante de seuils : $\epsilon_1..\epsilon_T$
Pour n allant de 1 au nombre de simulations souhaité N, faire :

1. A l'instant $t = 1$,
   Pour $i = 1, .., N$,
   Tant que $\rho(S(x), S(y)) < \epsilon_1$
   Simuler $\theta_i^{(1)} \sim \pi(\theta)$ et $x \sim p(x|\theta_i^{(1)})$
   $w_i \leftarrow 1/N$
   $\tau_2^2 \leftarrow 2 \times \big($ variance empirique des $\theta_i^{(1)}\big)$
2. A l'instant t, $\quad 2 \leq t \leq T$
   Pour $i = 1, .., N$,
   Tant que $\rho(S(x), S(y)) < \epsilon_t$
   Tirer aléatoirement θ_i^* des $\theta_j^{(t-1)}$, chacun affecté de la probabilité $w_j^{(t-1)}$
   Simuler $\theta_i^{(t)} \sim K(\theta|\theta_i^*; \tau_t^2)$ et $x \sim p(x|\theta_i^{(t)})$
   Choisir les poids $w_i \propto \pi(\theta_i^{(t)} / \sum_j = 1^N w_j^{(t-1)} K(\theta_i^{(t)}|\theta_j^{(t-1)}; \tau_t^2)$
   $\tau_{t+1}^2 \leftarrow 2 \times \big($ variance empirique pondérées des $\theta_i^{(t)}\big)$

3 Application en épidémiologie

3.1 Modèle SIR avec contact tracing

Nous étudierons dans cette partie l'application des méthodes ABC à l'épidémie du VIH à Cuba suivant l'article de Blum et Tran [2]. Le modèle est formé de trois variables : les individus sains S, les individus infectés I, et les individus repérés R. On suppose qu'une fois un individu repéré, il ne contamine plus personne. Il est alors soit dans la classe R_1 s'il a été repéré par un test aléatoire (prise de sang, grossesse,...), soit dans la classe R_2 s'il l'a été par contact tracing. Rappelons que le contact tracing est une méthode de pistage des individus HIV positifs : à chaque dépistage, on demande à un individu infecté qui ont été ses partenaires sexuels afin de les prévenir du risque encouru. On peut quitter où arriver dans une certaines classe au taux indiqués dans la figure 3.

On notera $\theta = (\mu_1, \lambda_1, \lambda_2, \lambda_3, c)$, qui paramètre le modèle. Il a été démontré par Clemençon *et al.* [3] que le système d'EDS obtenu à partir de ce modèle converge à la limite macroscopique (population de grande taille) vers les équations classiques de l'épidémiologie :

$$\begin{cases} s_t' &= \mu_0 - \lambda_0 s_t - \lambda_1 s_t i_t \\ i_t' &= \lambda_1 s_t i_t - (\mu_1 + \lambda_2) i_t - \lambda_3 i_t \int \varphi(a) \rho_t(a) da \\ \rho_t(0) &= \lambda_2 i_t + \lambda_3 i_t \int \varphi(a) \rho_t(a) da \end{cases}$$

(ici $\varphi(t) = e^{-ct}$)

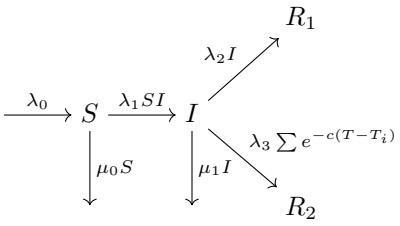

FIGURE 3 – Modèle SIR incorporant le contact-tracing.

Voici un algortihme de simulation du modèle tiré de l'article de Blum et Tran [2] :

1. Initialisation : $S = S_0$ individus sains, $I = I_0$ individus infectés, $R_i = 0$ individu détécté. Temps d'arrêt $T > 0$.
   Sachant que l'on a simulé k événements aus temps t_j pour $j = 1, k$, et que le temps courant $\tau = t_k$:
2. Simuler une exponentielle $\mathcal{E}(C_k)$ où
$$C_k = \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} + \lambda_3 I_{t_k} R_{t_k}.$$
3. Actualiser le temps courant $\tau \leftarrow \tau + \mathcal{E}$.
4. Si $\tau > T$, fin. Sinon : simuler une uniforme U sur $[0, C_k]$ et
   (a) Si $0 \leq U \leq \lambda_1 S_{t_k} I_{t_k}$: $S \leftarrow S - 1$ et $I \leftarrow I + 1$.
   (b) Si $\lambda_1 S_{t_k} I_{t_k} \leq U \leq \lambda_1 S_{t_k} I_{t_k} + \mu_1 I_{t_k}$: $I \leftarrow I - 1$.
   (c) Si $\lambda_1 S_{t_k} I_{t_k} + \mu_1 I_{t_k} \leq U \leq \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k}$: $I \leftarrow I - 1$ et $R_1 \leftarrow R_1 + 1$.
   (d) Si $\lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} \leq U \leq \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} + \lambda_3 I_{t_k} \sum e^{-c(\tau - T_i)}$: $I \leftarrow I - 1$, $R_2 \leftarrow R_2 + 1$
   (e) Sinon : ne rien faire.

La figure 4 illustre cet algorithme avec un code R disponible entièrement à la page web : https://github.com/cdellaie/ABCEpidemiolgy (fichier ABCandSIR.R). On a simulé 10000 événements avec une population de 201 individus, avec les paramètres $(\mu_1, \lambda_1, \lambda_2, \lambda_3) = (0.5, 0.5, 0.2, 0.3)$.

3.2 Méthodes ABC

Nous présentons dans cette section coment appliquer les méthodes ABC au modèle précédent, *SIR* avec *contact tracing*. Les paramètres du modèle prennent

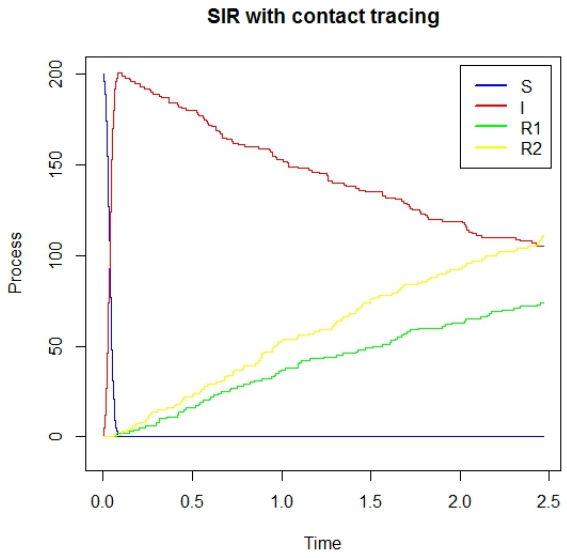

FIGURE 4 – Simulations

pour *prior* :

$$(\log \mu_1, \log \lambda_1, \log \lambda_2, \log \lambda_3) \quad \sim \quad U_{(-6,-4)} \otimes U_{(-9,-6)} \otimes U_{(-4,3)} \otimes U_{(-8,2)}$$

où $U_{(a,b)}$ désigne une loi uniforme sur l'intervalle (a, b).

On suppose que l'on dispose de données observées, à savoir le nombre d'individus détéctés chaque année $R^1_{obs}(t)$ et $R^2_{obs}(t)$. On simule alors N trajectoires R^j_l selon le *prior* ci-dessus, et on observe la norme $\mathcal{L}^1$ entre les données simulées et observées :

$$|R^j_{obs} - R^j_l|_1 = \int_0^T |R^j_{obs}(t) - R^j_l(t)|dt \quad , l = 1..N, j \in \{1, 2\}.$$

On choisit un noyau produit pour les poids : $W_l = K_{\delta_1}(|R^1_{obs} - R^1_l|_1)K_{\delta_2}(|R^2_{obs} - R^2_l|_1) \quad , l = 1, N.$

Nous aurions aimé pouvoir simuler sur des données réelles, mais par manque de temps, ce rapport s'arrête ici.

Références

[1] Beaumont. Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution and Systematics*, 41 :379–406, 2010.

[2] Viet Chi Tran Michael G.B. Blum. Approximate bayesian computation for epidemiological models : Applications to the cuban hiv-aids epidemic with contact tracing and unobserved infectious population. 2009.

[3] V.C. Tran S. Clemencon and H. De Arazoza. A stochastic sir model with contact tracing : large population limit with statistical applications. *Journal of Biological Dynamics*, 2008.