Inférence bayésienne de paramètres par forêts aléatoires de régression et ABC.

Louis Raynal [1] & Jean-Michel Marin [1]

[1] *Institut Montpelliérain Alexander Grothendieck, CNRS, Université de Montpellier*
louis.raynal@umontpellier.fr ; jean-michel.marin@umontpellier.fr

Résumé. Face à la complexité grandissante de certains modèles statistiques, la vraisemblance peut ne pas être disponible ou calculable, elle est alors dite inaccessible. Dans un contexte bayésien, des techniques de simulations intensives se sont développées, les méthodes de calcul bayésien approché (de l'anglais *Approximate Bayesian Computation* ou ABC) en font partie. Elles comparent des résumés statistiques de données observées et simulées, mais nécessitent cependant une calibration minutieuse, notamment dans leur choix. Nous proposons une nouvelle approche mélangeant ABC et forêts aléatoires de régression (Breiman, 2001), pour faire de l'inférence de paramètres. L'idée est d'utiliser une table de référence ABC comme échantillon d'apprentissage pour des forêts de régression, une par dimension de l'espace des paramètres, dans le but d'estimer espérances, variances ou quantiles *a posteriori*. La covariance entre paramètres peut être gérée par des forêts supplémentaires. Cette méthodologie a été ajoutée à la bibliothèque R `abcrf` et sera comparée aux résultats de techniques ABC standards. Tous les résultats de cette présentation sont davantage détaillés dans Marin et al. (2016).

Mots-clés. Calcul bayésien approché, inférence bayésienne, méthodes à vraisemblance inaccessible, forêts aléatoires.

Abstract. Since statistical models are getting increasingly complex, handling the likelihood function is becoming more and more of an issue. We now face situations where the likelihood cannot be computed or is simply unavailable, it is mentioned as intractable. In a Bayesian context, intensive simulation based methods have been developed : Approximate Bayesian Computation (ABC) is one of them. Observed and simulated summary statistics are compared, however calibration on their choice can be fastidious. We propose to conduct Bayesian inference by mixing ABC and the random forest methodology of Breiman (2001) when applied to regression. We advocate the derivation of a new random forest for each component of the parameter vector, trained on a so called ABC reference table, to derive posterior means, variances or quantiles. Covariance between parameter components are handled by separate random forests. The methodology introduced here has been added to the `abcrf` R library and will be compared with standard ABC solutions. All presented results are further detailed in Marin et al. (2016).

Keywords. Approximate Bayesian Computation, Bayesian inference, likelihood-free methods, random forests.

1 Introduction et contexte

Les modèles statistiques étant de plus en plus complexes, le calcul de la fonction de vraisemblance $f(y \mid \theta)$ peut être un problème. En effet, elle peut ne pas être exprimable comme fonction des paramètres, ou bien ne pas être calculable en un temps raisonnable, les techniques classiques basées sur celle-ci deviennent alors inutilisables. Une option pour passer outre cette difficulté réside dans les méthodes de simulation intensives dites *Approximate Bayesian Computation* (ABC) (Beaumont et al., 2002 ; Csilèry et al., 2010 ; Marin et al., 2012). Dans le cadre de celles-ci, bien que $f(y \mid \theta)$ soit incalculable, pour un paramètre θ fixé il est indispensable de savoir générer des observations selon le modèle.

Le principe de l'ABC est simple : comparer des données observées et simulées selon un paramètre θ lui-même généré grâce à une loi *a priori*. Selon leur proximité, ce θ est accepté ou rejeté comme une réalisation de la loi *a posteriori* du paramètre sachant l'observation. En pratique, des résumés statistiques des données sont utilisés pour mesurer la proximité. Deux difficultés se posent cependant : un nombre de simulations très important et une calibration minutieuse sont nécessaires, notamment dans le choix de résumés pertinents.

Un échantillon selon les lois *a posteriori* des paramètres résulte ainsi de l'ABC. Les praticiens ne sont bien souvent pas intéressés par la loi en elle-même, mais par certains moments d'intérêt, tels la moyenne, la variance ou des quantiles *a posteriori*. C'est pourquoi nous présentons une méthode les estimant directement. Celle-ci se base sur l'outil d'apprentissage statistique que sont les forêts aléatoires de régression (Breiman, 2001). La motivation principale de leur usage étant la robustesse dont elles font preuve en présence de covariables de bruit, dans un cadre parcimonieux. Ainsi utiliser ces forêts permettrait d'éviter, entre autres, une calibration en terme de choix de résumés, c'est pourquoi nous ne les sélectionnerons pas mais au contraire utiliserons tous ceux dont on dispose, supposés arbitrairement grands. Aucune sélection du nombre de régresseurs n'est donc effectuée. Toutes les méthodes et résultats présentés lors de cette présentation sont disponibles dans Marin et al. (2016).

2 Méthodologie

Nous considérons le modèle statistique suivant

$$\{f(y \mid \theta) \colon y \in \mathcal{Y}, \theta \in \Theta\}\,,\ \mathcal{Y} \subseteq \mathbb{R}^n\,,\ \Theta \subseteq \mathbb{R}^p\,, \qquad p, n \geq 1$$

et plaçons sur le paramètre θ une loi *a priori* $\pi(\theta)$. Pour un vecteur d'observations y^*, nous souhaitons faire de l'inférence bayésienne sur le paramètre θ^* dont il est issu. Comme mentionné ci-dessus, nous nous intéressons à l'estimation d'espérances, variances et de quantiles *a posteriori*. La difficulté principale étant le caractère incalculable de la vraisemblance $f(y \mid \theta)$.

Face à un tel problème, l'ABC utilise ce que l'on appelle une table de référence ABC. L'algorithme 1 présente sa construction. Des paramètres sont simulés suivant $\pi(\theta)$, puis de

nouvelles données sont générées selon le modèle en utilisant ces valeurs de paramètres. Ces y simulés sont ensuite résumés par une fonction $\eta\colon \mathcal{Y} \to \mathbb{R}^k$, nous travaillons ainsi avec k résumés statistiques. L'idée principale que nous proposons est l'utilisation de cette table comme échantillon d'apprentissage pour des forêts aléatoires de régression, la particularité étant que nous réaliserons une forêt par dimension de l'espace des paramètres θ. Nous avons intitulé cette stratégie, ODOF pour *One Dimension One Forest*.

Algorithme 1 : Simulation d'une table de référence de taille N.

pour $t \leftarrow 1$ **à** N **faire**
  Simuler $\theta^{(t)} \sim \pi(\theta)$;
  Simuler $y^{(t)} = (y_1^{(t)}, \ldots, y_n^{(t)}) \sim f\left(y \mid \theta^{(t)}\right)$;
  Calculer $\eta(y^{(t)}) = \{\eta_1(y^{(t)}), \ldots, \eta_k(y^{(t)})\}$;
fin

Si l'on s'intéresse à des informations sur θ_j, la $j^{\text{ème}}$ composante de θ, on construit alors une forêt de régression aléatoire de taille B sur les données d'entrainement $\{\theta_j^{(t)}, \eta(y^{(t)})\}$, pour $t = 1, \ldots, N$.

Espérance. Ainsi, pour un vecteur y^* la prédiction associée sera égale à la moyenne des prédictions fournies par chaque arbre. La forêt estime donc l'espérance *a posteriori* $\mathbb{E}\left(\theta_j \mid \eta(y^*)\right)$ par

$$\widetilde{\mathbb{E}}\left(\theta_j \mid \eta(y^*)\right) = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{\mid L_b(\eta(y^*)) \mid} \sum_{\{t:\eta(y^{(t)})\in L_b(\eta(y^*))\}} n_b^{(t)}\theta_j^{(t)}. \tag{1}$$

Le terme $L_b(\eta(y^*))$ désigne la feuille de l'arbre b où l'observation résumée $\eta(y^*)$ tombe après être passée à travers celui-ci, $\mid L_b(\eta(y^*)) \mid$ est égal au nombre d'individus dans cette feuille, $n_b^{(t)}$ le nombre de fois où l'observation $\{\theta_j^{(t)}, \eta(y^{(t)})\}$ est présente dans l'échantillon *bootstrap* ayant servi à la construction de l'arbre b. Notons que $n_b^{(t)} = 0$ indique que l'observation n'a pas été utilisée, elle porte alors l'appellation *out-of-bag* (OOB). Comme souligné par Meinshausen (2006), nous pouvons réécrire l'égalité (1) comme une moyenne pondérée des données d'apprentissage θ_j. Nous avons alors

$$\widetilde{\mathbb{E}}\left(\theta_j \mid \eta(y^*)\right) = \sum_{t=1}^{N} \omega_t(\eta(y^*))\theta_j^{(t)}. \tag{2}$$

Quantile. Ces poids jouent un rôle clé dans notre approche. En effet, nous pouvons les utiliser pour déterminer des quantiles *a posteriori* d'ordre α, au travers de l'approximation

de la fonction de répartition

$$\widetilde{\mathrm{F}}(x \mid \eta(y^*)) = \sum_{t=1}^{N} \omega_t(\eta(y^*)) \mathbf{1}_{\{\theta_j^{(t)} \leq x\}}. \tag{3}$$

Variance. Nous nous proposons de plus d'utiliser ces $\omega_t(\eta(y^*))$ pour pondérer le vecteur des erreurs *out-of-bag* issus de la forêt portées au carré, c'est-à-dire

$$\left\{ \left(\theta_j^{(t)} - \mathbb{E}_{\mathrm{OOB}}(\theta_j \mid \eta(y^{(t)}))\right)^2 \right\}_{t=1,\ldots,N},$$

dans le but d'estimer la variance *a posteriori* par

$$\widetilde{\mathbb{V}}(\theta_j \mid \eta(y^*)) = \sum_{t=1}^{N} \omega_t(\eta(y^*)) \left(\theta_j^{(t)} - \mathbb{E}_{\mathrm{OOB}}\left(\theta_j \mid \eta(y^{(t)})\right)\right)^2. \tag{4}$$

Nous présenterons d'autres variantes à l'estimation de variances *a posteriori*, utilisant la fonction de répartition approximée en (3) ou bien faisant intervenir une nouvelle forêt.

Covariance. La covariance entre deux paramètres θ_j et θ_ℓ peut être estimée grâce à une nouvelle forêt. Nous nous proposons de la construire en nous basant sur le produit des erreurs *out-of-bag* entre les deux paramètres.

La bibliothèque R `abcrf`, initialement créée pour le choix de modèle ABC par forêts aléatoires décrit dans Pudlo et al. (2015), comprend la majorité des méthodes présentées ici. La construction des forêts s'appuie sur le package R `ranger`, profitant ainsi de ses qualités de rapidité, de parallélisation et d'optimisation mémoire.

3 Étude de cas

Nous comparerons la qualité de notre approche par forêts au travers d'un exemple de régression simulé, où les distributions *a posteriori* des différents paramètres sont connues. Étant donnée une matrice de design $X = [x_1, x_2]$ de taille $n \times 2$, nous considérons le modèle hiérarchique

$$\begin{aligned}(y_1, \ldots, y_n) \mid \beta_1, \beta_2, \sigma^2 &\sim \mathcal{N}_n(X\beta, \sigma^2 Id),\\ \beta_1, \beta_2 \mid \sigma^2 &\sim \mathcal{N}_2(0_2, n\sigma^2 (X^\top X)^{-1}),\\ \sigma^2 &\sim I\mathcal{G}(4, 3),\end{aligned}$$

où $\beta = (\beta_1, \beta_2)^\top$. $\mathcal{N}_m(\mu, \Sigma)$ dénote la loi normale multidimensionnelle de dimension m de vecteur moyen μ et de matrice de variance-covariance Σ. $I\mathcal{G}(\kappa, \lambda)$ désigne la loi inverse-gamma avec pour paramètre de forme κ et d'échelle λ. Ainsi, sous la condition que $X^\top X$ est inversible nous disposons des lois *a posteriori* explicites pour β_1, β_2 et σ^2.

	ODOF	ARR	ANN
$\mathbb{E}(\beta_1 \mid y)$	**0.09**	0.12	0.14
$\mathbb{E}(\beta_2 \mid y)$	**0.11**	0.26	0.27
$\mathbb{E}(\sigma^2 \mid y)$	**0.04**	0.06	0.07
$\mathbb{V}(\beta_1 \mid y)$	**0.50**	0.88	0.61
$\mathbb{V}(\beta_2 \mid y)$	**0.46**	0.89	0.61
$\mathbb{V}(\sigma^2 \mid y)$	**0.31**	0.90	0.73
$\mathbb{C}ov(\beta_1, \beta_2 \mid y)$	**0.26**	0.85	0.64
$\mathbb{Q}_{0.025}(\beta_1 \mid y)$	0.29	0.37	**0.27**
$\mathbb{Q}_{0.025}(\beta_2 \mid y)$	0.31	0.40	**0.25**
$\mathbb{Q}_{0.025}(\sigma^2 \mid y)$	**0.05**	0.22	0.18
$\mathbb{Q}_{0.975}(\beta_1 \mid y)$	**0.43**	0.79	0.81
$\mathbb{Q}_{0.975}(\beta_2 \mid y)$	**0.47**	0.86	0.55
$\mathbb{Q}_{0.975}(\sigma^2 \mid y)$	**0.10**	0.12	0.12

TABLE 1 – Comparaison des erreurs absolues moyennes normalisées (NMAE) pour les estimations de quantités d'intérêt par forêts (ODOF), régression Ridge ajustée (ARR) et méthode *neural network* ajustée (ANN).

Nous utilisons une table de référence de taille $N = 10000$, des échantillons de taille $n = 100$, $k = 60$ résumés statistiques : les estimations par maximum de vraisemblance de β_1, β_2, la somme des carrés résiduels, la covariance et corrélation empirique entre y et x_1, la covariance et corrélation empirique entre y et x_2, la moyenne, variance et médiane de notre échantillon et enfin 50 variables de bruit indépendantes simulées selon une loi uniforme $\mathcal{U}_{[0,1]}$. Ce bruit a pour but de se placer dans un contexte parcimonieux, où seul un faible nombre de covariables est important. Une table de test de taille $N_{\text{pred}} = 100$ est utilisée afin de confronter la qualité de nos estimateurs avec les méthodes standards de l'ABC, notamment celles avec ajustement selon des régressions Ridge ou encore des réseaux de neurones, (Blum et al. 2013 ; Beaumont, 2010).

Au travers de tableaux d'erreurs et de résultats graphiques, nous jugerons de la performance de la méthodologie précédemment introduite contre l'état de l'art ABC. La table 1 résume les erreurs absolues moyennes normalisées (NMAE) pour les différentes estimations par l'approche ODOF et les deux méthodes ABC avec ajustement mentionnées ci-dessus. Les résultats obtenus sont très encourageants, notamment dans un cas parcimonieux, les forêts aléatoires permettant par ailleurs de passer outre le grand aspect calibration présent dans les techniques ABC actuelles. Nous pensons qu'une telle association entre ABC et *machine learning* peut ouvrir un large champ de possibilités en terme d'estimation de paramètres lorsque la vraisemblance est inaccessible.

Bibliographie

[1] Beaumont, M., Zhang, W. et Balding, D. (2002), Approximate Bayesien computation in population genetics, *Genetics*, 162 :2025-2035.

[2] Beaumont, M. (2010), Approximate Bayesian computation in evolution and ecology, *Annual Review of Ecology, Evolution, and Systematics*, 41 :379-406.

[3] Blum, M., Nunes, M., Prangle, D. et Sisson, S. (2013), A comparative review of dimension reduction methods in Approximate Bayesian Computation, *Statistical Science*, 28(2) :189-208.

[4] Breiman, L. (2001), Random forests, *Machine Learnings*, 45(1) :5-32.

[5] Csillèry, K, Blum, M., Gaggiotti, O. et François, O (2010), Approximate Bayesian computation (ABC) in practice, *Trends in Ecology and Evolution*, 25 :410-418.

[6] Marin, J.-M., Pudlo, P., Robert, C. P.. et Ryder, R. (2012), Approximate Bayesian computation methods, *Statistics and Computating*, 22(6) :1167-1180.

[7] Marin, J.-M., Raynal, L., Pudlo, P., Ribatet, M. and Robert, C. P. (2016), ABC random forests for Bayesian parameter inference, *ArXiv e-prints*, 1605.05537.

[8] Meinshausen, N. (2006), Quantile Regression Forests, *Journal of Machine Learning Research*, 7 :983-999.

[9] Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M. et Robert, C. P. (2015), Reliable ABC model choice via random forests, *Bioinformatics*, 32(6) :859-866.