ABC pour le choix de modèle de formation stellaire des galaxies

Grégoire Aufort [1,2] & Pierre Pudlo [1] & Laure Ciesla [2] & Véronique Buat [2]

[1] *Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France*
[2] *Aix Marseille Univ, CNRS, CNES, LAM, Marseille, France*
gregoire.aufort@univ-amu.fr / pierre.pudlo@univ-amu.fr / laure.ciesla@lam.fr / veronique.buat@lam.fr

Résumé. Nous nous intéressons au problème du choix de modèle bayésien dans le cas où un nombre important de jeux de données, ou d'objets doivent être traités. Nous proposons une extension de l'algorithme ABC-RandomForest pour le choix de modèle, basée sur du boosting d'arbres (minimisation de l'entropie croisée) sur le catalogue de simulations ABC. Cet algorithme d'apprentissage nous permet de contourner l'emploi de statistiques résumées dans notre algorithme ABC. Nous présentons une application à l'astrophysique. À partir de données photométriques, nous montrons la pertinence de la complexification d'un modèle d'histoire de formation stellaires pour une proportion non négligeable de jeux de données parmi des dizaines de milliers de galaxies.

Mots-clés. ABC, Choix de modèle bayésien, Boosting, Histoire de formation stellaire

We are interested in the bayesian model choice problem when a large number of objects have to be processed. We propose an extension of the ABC-RandomForest algorithm for model choice, based on crossentropy minimization on the ABC simulation catalog. This learning algorithm allows us to bypass the use of summary statistics for ABC. We present an application in astrophysics. From photometric data, we show the relevance of the complexification of a stellar formation history model for an important part of the datasets among tens of thousands of galaxies.

Keywords. ABC, bayesian model choice, Boosting Star formation history

L'estimation de l'histoire de formation stellaire d'une galaxie est un problème important en astrophysique. Cette histoire de formation stellaire (SFH) est une courbe paramétrée, modélisant l'évolution de la masse stellaire formée au cours de temps. Elle laisse une trace dans la distribution spectrale d'énergie (SED) émise par la galaxie, que l'on observe avec des données photométriques à différentes longueurs d'ondes (ici, 14). Un modèle physique à 17 paramètres (complémentaires aux paramètres de SFH) permet de relier l'histoire de formation stellaire à la distribution d'énergie spectrale, et de tenir compte du bruit de mesure.

Différentes formes de parmétrisations de la SFH ont été proposées dans la littérature (voir Ciesla *et al.*, 2017), puisque toutes les galaxies n'ont pas la même histoire. Notre

objectif ici est de choisir une paramétrisation (un modèle statistique) par galaxie observée. Nous allons donc conduire une analyse bayésienne de choix de modèles pour chacun des jeux de données. Ici, nous nous concentrons sur l'analyse de 42005 galaxies, qui représentent autant de jeux de données observées.

Nous proposons une méthode de choix de modèles par calcul bayésien approché (*Approximate Bayesian Computation* ou ABC, voir Marin *et al.* 2012). Les techniques classiques de calcul du facteur de Bayes (par exemple Robert et Casella, 2013), pourraient être employées ici. Mais il faut alors mettre en œuvre un algorithme de Monte–Carlo par jeu de données à analyser. Ce n'est pas le cas des algorithmes ABC non adaptatifs. En effet, en utilisant uniquement des tirages suivant la loi a priori (qui est commune à toutes les galaxies), nous obtenons ainsi un algorithme efficace, qui est peu sensible au nombre de jeux de données à analyser, donc de galaxies observées. À la suite de Pudlo *et al.* (2016), nous proposons ici un algorithme ABC pour le choix de modèle :

- qui utilise des techniques de machine learning pour apprendre sur les simulations, comme ABC-RF,
- mais fournit une approximation de la probabilité a posteriori en une seule étape, contrairement à ABC-RF.

Au travers de cette application en astrophysique, nous souhaitons souligner un avantage de cet algorithme ABC : la possibilité de factoriser la majorité des calculs lors de l'analyse d'un grand nombre de jeux de données. En effet, la simulation des jeux de données suivant la loi a priori et l'entraînement de l'algorithme de machine learning sont communs à l'analyse de toutes les galaxies. La première partie de ce résumé détaille l'algorithme ABC, et la seconde partie présente les résultats astrophysiques. Les avantages de notre proposition sont donnés en conclusion.

1 Calcul bayésien approché pour le choix de modèle

Les méthodes Approximate Bayesian Computation (ABC) pour le choix de modèle se basent sur l'utilisation de données simulées d'après chacun des M modèles pour un nombre important de pseudo-données x selon des valeurs θ_m du paramètre échantillonées suivant la loi a priori $\pi_m(\theta_m)$ du m-ème modèle. Ces simulations forment une table de référence dont les éléments sont comparés à l'observation x_0. On notera (m_i, θ_i, x_i), $i = 1, \ldots, N$, l'ensemble des simulations de la table de référence. L'objectif est de les comparer au jeu de données observée x_0, via une fonction de résumé statistique S, pour approcher les probabilités a posteriori

$$\pi(m|x_0) \propto \int \mathrm{d}\theta_m \pi(m)\pi_m(\theta) f(x_0|\theta, m), \tag{1}$$

où $\pi(m)$ est la probabilité a priori du modèle numéro m, $\pi_m(\theta)$ la densité de probabilité a priori en θ pour le modèle m, et $f(\cdot|\theta, m)$ la densité engendrant les données dans le modèle m.

L'algorithme ABC naïf estime la probabilité a posteriori du modèle m par la proportion de simulations (m_i, θ_i, x_i) issues de ce modèle lorsque que l'on ne garde que celles qui sont les plus proches de l'observation (au sens d'une distance d entre les statistiques résumées $S(x)$ et $S(x_0)$). Ainsi, cet algorithme produit un échantillon distribué selon la loi jointe

$$\pi_\epsilon(m, \theta|S(x_0)) = \frac{\pi(m)\pi_m(\theta)f(x|\theta, m)\mathbf{1}_{B(x_0,\epsilon)}(x)}{\sum_{m'} \pi(m') \int_{B(x_0,\epsilon)\times\Theta} \pi_{m'}(\theta)f(x|\theta, m')\mathrm{d}x\mathrm{d}\theta}, \tag{2}$$

où $B(x_0, \epsilon) = \{x : d(S(x), S(x_0)) \leq \epsilon\}$. Ce qui permet d'estimer $\pi(m|x_0)$ par marginalisation triviale.

Lorsque le seuil ϵ est calibré pour que le nombre de simulations tombant dans $B(x_o, \epsilon)$ soit égal à k, cet algorithme naïf est équivalent à la méthode des k plus proches voisins. Notons aussi que, lorsque $\epsilon \to 0$, la $\pi_\epsilon(m|S(x_0))$ converge vers

$$\pi(m|S(x_0)) \propto \pi(m) \int_{B(x_0,0)\times\Theta} \pi_m(\theta)f(x|\theta, m)\mathrm{d}x\mathrm{d}\theta. \tag{3}$$

Le meilleur algorithme de classification prédisant m à l'aide des covariables du vecteur $S(x)$ au sens de l'erreur de classification est basé sur le classifieur de Bayes, donc la probabilité a posteriori (3). Celui-ci n'est pas atteignable dans le contexte des méthodes ABC, et doit donc être approché par un algorithme de classification entraîné sur la table de référence. Pudlo et al (2016) ont proposé d'utiliser Random Forest, un algorithme de classification moins sensible au fléau de la dimension que les k-plus proches voisins. Cette approche permet de ne pas avoir à limiter trop la dimension de la statistique résumée. Elle permet également une estimation extrêmement rapide une fois l'algorithme entraîné. Mais Random Forest ne fournit pas d'approximation de (3). Pudlo et al. (2016) ont donc proposé de retrouver cette probabilité en ajoutant une seconde forêt aléatoire.

En prolongeant l'idée d'une classification basée sur la table de référence, nous proposons une modification de l'algorithme ABC-RandomForest remplaçant la classification RandomForest par un algorithme tel que le Boosting d'arbre ou les réseaux de neurones, qui reposent sur le calcul de (3) avant de prendre la décision de classification (Richard et Lippman, 1991). Cette modification permet la suppression de la seconde étape de l'algorithme ABC-RandomForest, ainsi qu'une plus grande souplesse dans le choix de l'algorithme selon les spécificités du problème considéré : structure des données, taille de la table de référence... Nous explorons dans la suite la régression logistique, les réseaux de neurones à une 1 ou 3 couches, ainsi que le boosting d'arbres implémenté dans XGBoost (Chen et Guestrin, 2016), et choisissons le meilleur algorithme en terme d'erreur de classification sur la table de référence.

2 Choix de modèle de formation des galaxies

Nous nous intéressons à l'histoire de formation des galaxies. Ces galaxies sont observées sous la forme de flux d'énergie mesurés à différentes longueurs d'onde fixes, rangé dans un vecteur x_0.
Avec le logiciel CIGALE (Boquien *et al*, 2019), nous pouvons simuler une distribution spectrale d'énergie (SED) en convoluant un modèle de population stellaire par l'histoire de Formation stellaire, puis en appliquant l'effet de la poussière interstellaire sur la lumière observée.
Différents modèles physiques ont été proposés pour représenter l'histoire de formation stellaire des galaxies, paramétrés par θ. Nous nous intéressons ici au modèle Delayed, qui représente une phase de croissance de la production stellaire suivie d'une décroissance exponentielle. Nous nous interrogeons sur la nécessité d'ajouter à ce modèle un paramètre de flexibilité dans l'histoire récente des galaxies. Cette flexibilité peut être soit une augmentation soudaine de l'activité de production de la galaxie, soit au contraire une diminution.

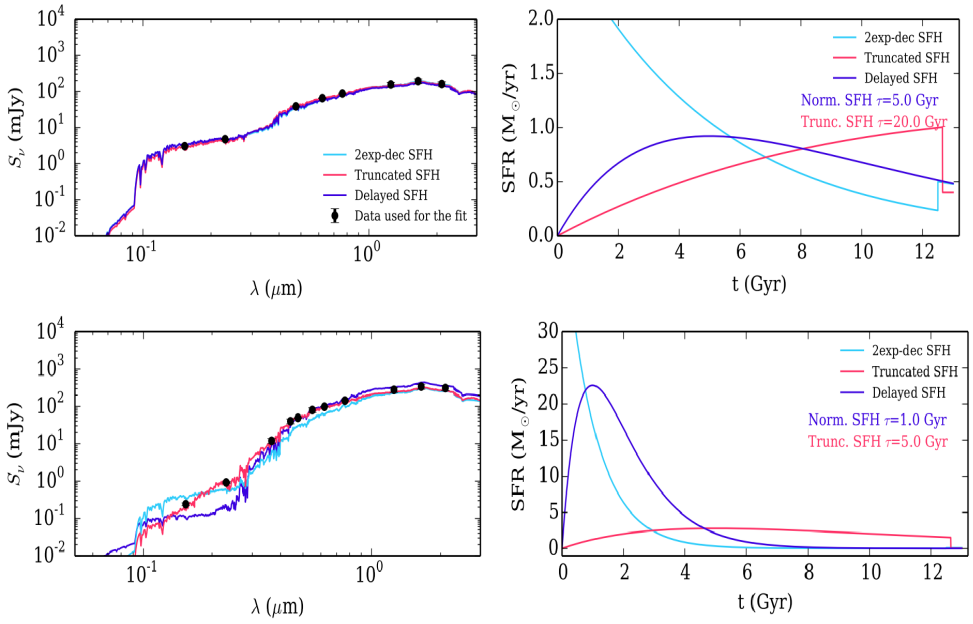

Figure 1: Deux exemples de trois SEDs (à gauche) issues de trois différentes SFH (à droite). Le modèle "Delayed + flex" (en rouge) est clairement plus proche des points mesurés pour l'exemple du bas. Aucune différence significative n'est observable pour les SEDs de l'exemple du haut, malgré des SFH très différentes

Table 1: Taux d'erreur de classification de chaque algorithme sur un ensemble de test simulé suivant les lois a priori

Algorithme	**Taux d'erreur (%)**	**Algorithme**	**Taux d'erreur (%)**
Logistic regression	30.27	1-layer-NN	22.51
LDA	30.43	3-layer-NN	21.06
k-nn (ABC naïf)	23.79	XGBoost	20.98

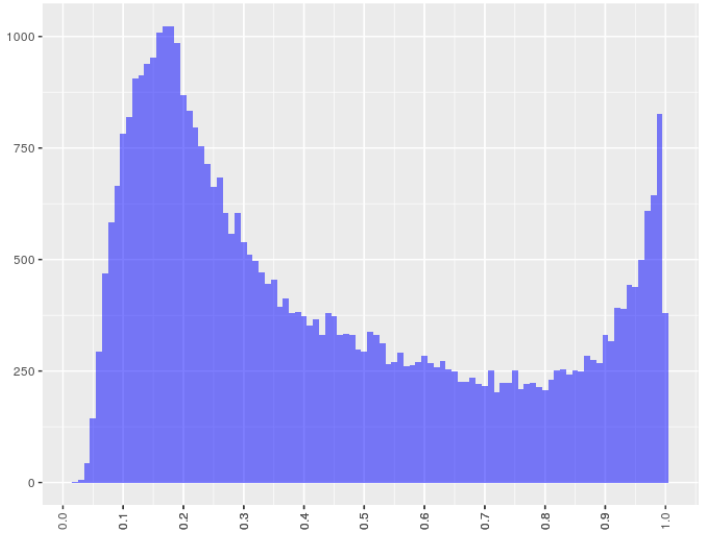

Figure 2: Distribution des estimations des probabilités a posteriori du modèle le plus complexe ("flex"), pour les 42005 galaxies.

Les 42005 SEDs observées sont issues de mesures sur autant de galaxies observées par le télescope spatial Hubble. Nous voudrions effectuer notre choix de modèle sur chacun de ces 42005 objets observés x_0. Pour cela, nous simulons une table de référence de 2.10^6 simulations (m_i, θ_i, x_i), issues d'une distribution a priori physiquement motivée et faiblement informative $\pi(m)\pi_m(\theta)$, pour chacun des deux modèles $m = 0$ (Delayed) ou 1 (Flex). En outre, vue la dimension modérée des données observées x_0, égale ici à 27, nous nous passons de l'utilisation de statistiques résumées, autrement dit $S(x) = x$. Ainsi, la probabilité (3) est égale à (1). Nous comparons plusieurs méthodes de classification sur ces données simulées, voir Table 1. Le temps de calcul pour nécessaire à l'estimation sur nos 42005 galaxies est de 56.8 minutes pour la méthode ABC naïve, contre 0.83 secondes pour la classification par boosting, implimentée dans XGboost. Les 42005 résultats sont résumés dans la figure 2 qui représente la distribution des estimations des 42005 probabilités a posteriori du modèle le plus complexe.

3 Conclusions

En étendant l'idée d'ABC-RandomForest, nous avons proposé une méthode de choix de modèle d'histoire de formation stellaire des galaxies. L'une des originalité de ce problÃ¨me est que le nombre de jeux de données observés est grand (42005 galaxies). Notre proposition présente plusieurs avantages :

- l'utilisation d'algorithmes robustes au fléau de la dimension, adaptés à la structure des données,
- la réduction drastique du temps de calcul au moment de l'estimation pour un grand nombre de jeux de données par rapport à un algorithme ABC naïf,
- un résultat plus précis, l'erreur d'approximation de (3) par XGBoost étant plus faible que celle des autres algorithmes,
- et toujours la non-utilisation d'une vraisemblance explicite.

Du point de vue astrophysique, nous avons mis en évidence la pertinence de l'ajout d'une composante de flexibilité dans l'histoire récente de formation stellaire pour la modélisation d'une partie conséquente du catalogue de galaxies étudiées.

Bibliographie

Boquien, M. Burgarella, D. Roehlly, Y. Buat, V. Ciesla, L. Corre, D. Inoue, A. K. et Salas, H (2019).,CIGALE: a python Code Investigating GALaxy Emission,*Astronomy & Astrophysics*, 622, A102

Chen, T. et Guestrin, G. (2016), XGBoost : A Scalable Tree Boosting System . *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794

Ciesla, L. Elbaz, D. et Fensch, J. (2017) The SFR-M main sequence archetypal star-formation history and analytical models. *Astronomy & Astrophysics* 608, A41.

Marin, J.M. Pudlo, P. Robert, C.P et Ryder, R.J (2012), Approximate Bayesian computational methods, *Statistics and Computing* 22, pp. 1167–1180

Pudlo, P. Marin, J-M. Estoup, A. Cornuet, J-M. Gautier, M et Robert, C.P. (2016) Reliable ABC model choice via random forests, *Bioinformatics*, 32, 6, pp. 859–866

Robert, C., et Casella, G. (2013). *Monte Carlo statistical methods*. Springer.

Richard, M.D. et Lippmann, R.P (1991), Neural network classiers estimate bayesian a posteriori probabilities, *Neural Computation* ,3, pp. 461–483