Mélange de processus Gaussiens multi-tâches et prédictions cluster-spécifiques

Arthur Leroy [1] & Pierre Latouche [2] & Benjamin Guedj [3] & Servane Gey [4]

[1] *Université de Paris, CNRS, MAP5 UMR 8145, F-75006 Paris, France et arthur.leroy.pro@gmail.com*

[2] *Université de Paris, CNRS, MAP5 UMR 8145, F-75006 Paris, France et pierre.latouche@u-paris.fr*

[3] *Inria, France et University College London, United Kingdom et benjamin.guedj@inria.fr*

[4] *Université de Paris, CNRS, MAP5 UMR 8145, F-75006 Paris, France et servane.gey@u-paris.fr*

Résumé. La communication proposée porte sur l'analyse de données fonctionnelles et la définition de modèles de processus Gaussiens (GP) multi-tâches pour traiter simultanément les problèmes de régression et de classification non-supervisée. L'algorithme Magma, issu d'un travail antérieur, permet la modélisation de multiples séries temporelles asynchrones supposées partager de l'information commune, offrant une amélioration drastique des performances comparées à la régression GP traditionnelle, ainsi qu'une pleine prise en compte de l'incertitude. Nous introduisons une extension de ce modèle permettant la définition d'un mélange de GPs multi-tâches ajoutant un aspect clustering à cette approche. L'apprentissage des hyper-paramètres d'un tel modèle repose sur la définition de distributions variationnelles, la vraisemblance n'étant pas calculable directement, permettant de conserver une formulation explicite des lois a posteriori. Des formules analytiques sont également dérivées pour la prédiction de temps non-observés. L'algorithme MagmaClust ainsi obtenu permet à la fois d'identifier des structures de groupes dans un ensemble de courbes et offre des prédictions cluster-spécifique encore améliorées par rapport à Magma. Cette approche a été implémentée et expérimentée sur différents jeux de données simulés, offrant des performances remarquables tant sur les aspects de clustering que de prédiction. Une application sur données réelles est également proposée via l'étude et la prédiction future des courbes de performances de jeunes nageurs français.

Mots-clés. Processus Gaussiens, apprentissage multi-tâche, clustering de courbes, méthodes variationnelles

Abstract. The present work is dedicated to the analysis of functional data and the definition of multi-task Gaussian processes (GP) models for simultaneously dealing with regression and clustering. The algorithm Magma, from a previous work, enables modelling multiple asynchronous time series, assumed to share information, offering a remarkable improvement in performances compared to standard GP regression, along

with a thorough quantification of uncertainty. An extension of this work is proposed from the definition of a multi-task GPs mixture, which enriches the previous approach with a clustering aspect. Learning the hyper-parameters in such model lies on the definition of variational distributions, since likelihood is not available directly, allowing us to maintain explicit posterior distributions. In addition, analytical formulas are derived for prediction of unobserved timestamps. The resulting algorithm, MagmaClust, offers a group-structure identification within a set of curves as well as enhanced predictions compared to Magma. This approach has been implemented and tested on several simulated datasets, exhibiting noticeable performances both on clustering and prediction tasks. A real data application, focusing on the study and forecast of future performance curves for young french swimmers, is proposed as well.

Keywords. Gaussian Processes, multi-task learning, curve clustering, variational inference

1 Contexte

Le cadre des processus Gaussiens (Rasmussen and Williams, 2006) offre une modélisation élégante pour traiter le cas des données fonctionnelles, mais souffre toutefois de limitations lorsque les points d'observations sont peu nombreux et/ou mal répartis sur le domaine d'étude. Cependant, la définition de modèles multi-tâches (Caruana, 1997), autorisant le partage d'informations, permet de tirer le meilleur parti de situations où de multiples séries temporelles, présentant des caractéristiques communes, sont observées. Une approche classique pour définir un modèle de GPs multi-tâche a été introduit dans Bonilla et al. (2008) en définissant une structure de covariance particulière, composée de deux matrices, représentant respectivement les covariances entre les individus et entre les tâches. Toutefois, tant sur le point de la complexité algorithmique, de l'impossibilité de gérer des observations asynchrones, que sur des capacités prédictives raisonnablement limitées, cette méthode reste non optimale dans de nombreuses applications. Plus récemment, un algorithme du nom de Magma a été proposé (Leroy et al., 2020b) pour traiter l'entrainement et la prédiction d'une nouvelle formulation de modèles de GPs multi-tâches. L'originalité de cette approche repose sur l'introduction d'un processus moyen, commun à tous les individus, qui, une fois estimé, embarque une information partagée fournissant une moyenne a priori pré-entrainée avant même la prédiction. Les performances prédictives se trouvent être grandement améliorées, notamment loin des points d'observations, tout en conservant une complète quantification de l'incertitude et une gestion naturelle des données observées irrégulièrement d'une courbe à l'autre.

2 Modèle et inférence

Le travail ici présenté (Leroy et al., 2020a) s'inscrit dans la continuité de cette approche, en proposant une généralisation du modèle précédent à l'aide d'un mélange de GPs, permettant d'identifier une éventuelle structure de groupe dans les multiples tâches d'entrainement. En effet, pour certains jeux de données, l'hypothèse d'un unique processus central sous-jacent peut être trop restrictive. Ainsi, pour une donnée fonctionnelle y_i associée au i-ème individu appartenant au k-ième groupe, le modèle génératif se définit comme suit :

$$y_i = \mu_k + f_i + \epsilon_i,$$

où μ_k est un GP spécifique au k-ème groupe, alors que f_i et ϵ_i représentent un GP et un bruit gaussien, tous deux spécifiques à l'individu i. Une formulation hiérarchique équivalente, comme donnée ci-dessous pour tout vecteur de temps d'observation $\mathbf{t}$, permet de mieux comprendre en quoi les processus μ_k définissent les moyennes de chacun des clusters:

$$p\left(y_i(\mathbf{t}) \mid \mu_k(\mathbf{t})\right) = \mathcal{N}\left(y_i(\mathbf{t});\ \mu_k(\mathbf{t}), \mathbf{\Psi}_{\theta_i,\sigma_i^2}(\mathbf{t},\mathbf{t})\right),\ \forall i, \forall \mathbf{t},$$

où $\mathbf{\Psi}_{\theta_i,\sigma_i^2}$ désigne la structure de covariance associée à l'individu i. Ce nouveau modèle dépend également d'une variable multinomiale latente Z_i, contrôlant l'appartenance des individus à chaque cluster. Dans cette approche, il est à présent nécessaire d'estimer les hyper-paramètres des noyaux de covariance, conjointement des lois hyper-posterior des processus μ_k et des variables Z_i. Les dépendances a posteriori entre ces dernières quantités poussent à introduire un algorithme Variationnel EM (VEM) (Attias, 2000) pour l'entrainement, où les hyper-paramètres sont obtenus par maximisation de l'ELBO via l'algorithme d'optimisation *L-BFGS-B* (Morales and Nocedal, 2011). Nous dérivons des lois variationnelles approximées, dont les expressions analytique permettent leur utilisation ultérieure dans de formules de prédiction GP. Un algorithme EM est également établi pour estimer les hyper-paramètres associés à un nouvel individu, partiellement observé, ainsi que ses probabilités d'appartenance aux différents clusters. Par intégrations successives sur les processus moyens μ_k, puis sur les Z_i, une loi a posteriori de mélange gaussien multi-tâche peut être déduite, définie comme une somme pondérée de prédictions GP cluster-spécifiques.

3 Expériences et résultats

Nous illustrons au travers de simulations les avantages d'une telle approche et son intérêt lorsque les données présentent des structures de groupes. Par exemple, la Figure 2 propose une comparaison sur un même jeu de données entre la régression GP classique, l'algorithme MAGMA, et notre nouvel approche MAGMACLUST, pour prédire des points non observés (rouge) à partir d'observations (noir), aidé par les données issues des individus d'entrainement (colorés en arrière plan). Les performances sur les aspects de clus-

tering sont évaluées sur la Figure 1, et celles-ci dépassent nettement celles d'alternatives usuelles de la littérature. L'algorithme a également été appliqué dans le cadre d'une étude des courbes de progressions de jeunes nageurs, issues de données de la fédération française de natation. Ce travail a permis d'indentifier différents profils de progression parmi les individus ainsi qu'une prédiction probabiliste fiable des performances futures pour chaque sportif.

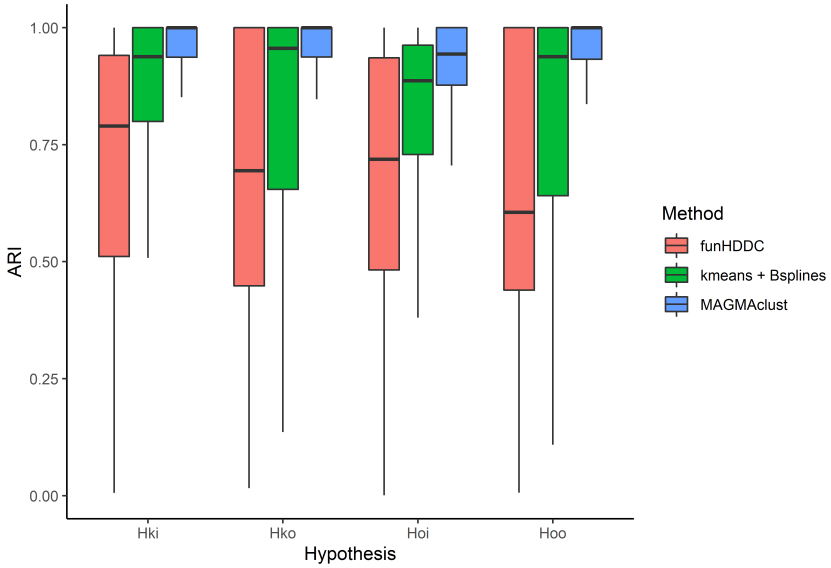

Figure 1: Valeurs des Adjusted Rand Index (ARI) entre les vrais clusters et les partitions estimées par les algorithmes kmeans, funHDDC, et MagmaClust. Le vrai nombre de groupes est spécifié dans chaque méthode et le ARI est calculé sur 100 jeux de données simulés selon 4 hypothèses différentes.

Bibliographie

Bibliographie

H. Attias. A Variational Baysian Framework for Graphical Models. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

E. V. Bonilla, K. M. Chai, and C. Williams. Multi-task Gaussian Process Prediction. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 153–160. Curran Associates, Inc., 2008.

R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, July 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734.

A. Leroy, P. Latouche, B. Guedj, and S. Gey. Cluster-Specific Predictions with Multi-Task Gaussian Processes. *PREPRINT arXiv:2011.07866 [cs, LG]*, Nov. 2020a.

A. Leroy, P. Latouche, B. Guedj, and S. Gey. MAGMA: Inference and Prediction with Multi-Task Gaussian Processes. *PREPRINT arXiv:2007.10731 [cs, stat]*, July 2020b.

J. L. Morales and J. Nocedal. Remark on algorithm L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 38(1):7:1–7:4, Dec. 2011. ISSN 0098-3500. doi: 10.1145/2049662.2049669.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass, 2006. ISBN 978-0-262-18253-9.

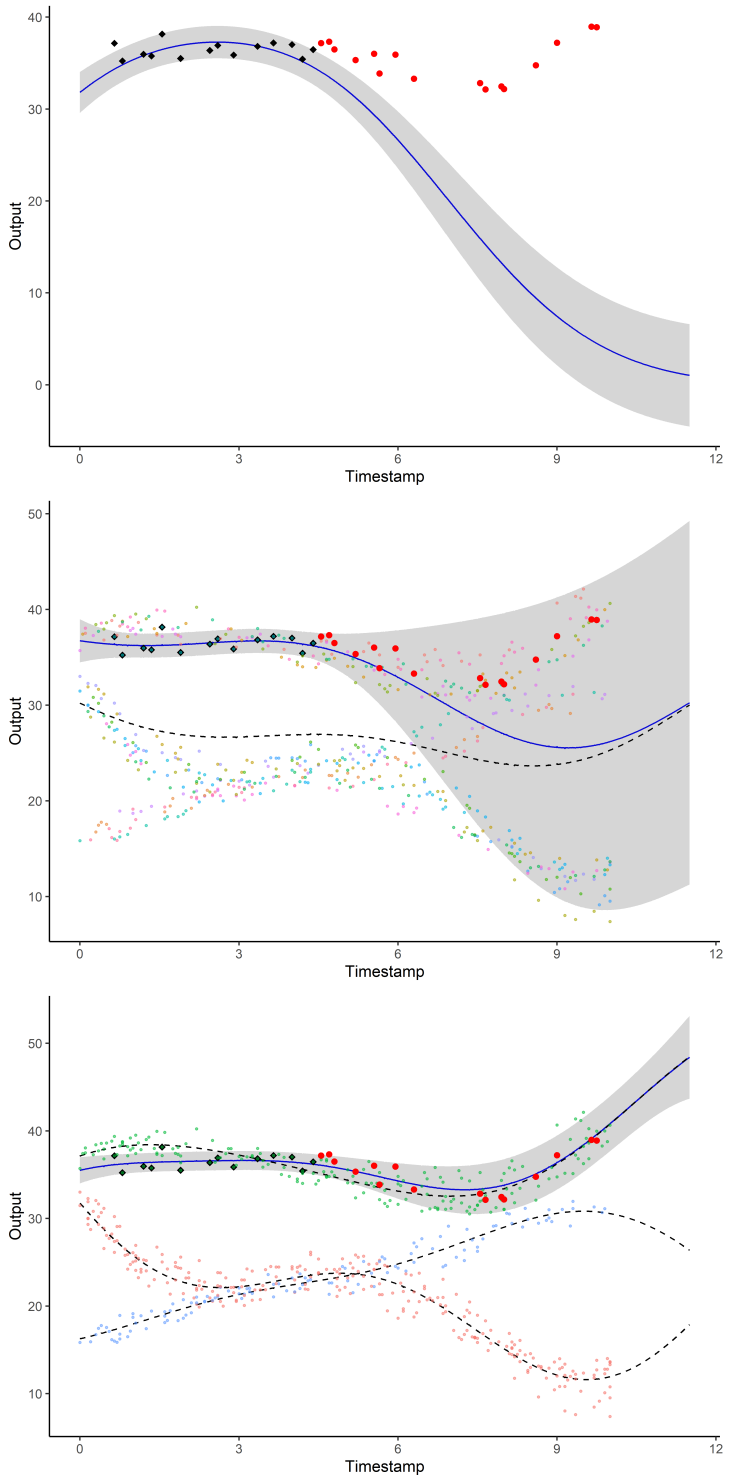

Figure 2: Courbes prédictives (bleu) et intervalles de crédibilité à 95% associés (gris) pour la régression GP (haut), MAGMA (milieu) et MAGMACLUST (bas). Les lignes pointillées représentent le pamarètre de moyenne de chaque processus moyens μ_k. Les points observés sont en noir, les points de test à prédire sont en rouge. Les points colorés en arrière plan sont issus des individus de la base d'entrainement.