Clustering et régression de données fonctionnelles par processus Gaussiens : Application à la prédiction de performance en natation

Arthur LEROY[*]
Servane GEY[†]
Pierre LATOUCHE[‡]

Résumé : Une grande part des données récoltées en science du sport vient de phénomènes dépendant du temps. Récemment, plusieurs structures sportives, comme les clubs ou les fédérations, ont collecté des données longitudinales dans l'espoir qu'elles puissent aider à la détection des jeunes à haut potentiel. Cependant, plusieurs études ont mis en avant le fait que la plupart des meilleurs jeunes ne restent pas au même niveau de performance une fois adulte. C'est pourquoi le problème de la détection pourrait bénéficier de méthodes d'analyse de données objectives et notamment du domaine de l'apprentissage statistique. Lors de cette étude, l'objectif réside dans la prédiction de performances futures d'un athlète à partir de ses performances passées et de l'information apportées par apprentissage sur les autres athlètes. La progression des sportifs étant intrinsèquement continue et les temps d'observations étant très irréguliers, les données seront considérées comme fonctionnelles et lissées à l'aide de fonctions de bases B-splines. Ces observations fonctionnelles sont supposées être des réalisations de processus Gaussiens, et le problème de prédiction est également traité par régression par processus Gaussien. Plus précisément, un modèle mixte est utilisé avec un processus moyen commun à tous les individus sommé à un processus d'effets aléatoires individuels. Cette approche permet d'utiliser l'information de tous les individus pour la modélisation et règle ainsi le problème du faible nombre d'observations irrégulières. Préalablement, une étape de clustering est appliquée sur les données fonctionnelles, permettant par la suite une prédiction dépendante du groupe d'appartenance d'un individu. La procédure est estimée par une approche Bayésienne, qui permet de prendre en compte l'incertitude de modélisation et de prédiction naturellement, ainsi que le calcul d'intervalles de crédibilité. Une étude sur des simulations sera présentée ainsi que l'application sur un jeu de données réelles provenant de la Fédération Française de Natation. L'intérêt de ce travail est double, offrant une meilleure compréhension du phénomène de progression dans le sport, et fournissant un outil d'aide à la décision pour la détection de jeunes talents.

Mots clefs. Données fonctionnelles, Processus Gaussiens, Prédiction, Sport haut niveau, B-splines, ...

Abstract. Many data collected in sport science come from time dependent phenomena. Several sport structures such as federations or clubs have collected these longitudinal data, hoping it may help the identification of future talents. Moreover, recent studies highlighted that most of top-level young athletes do not remain at the best level of performance at adult age. Thus, the talent identification issue might benefit from data analysis and particularly from machine learning methods. In the present paper, the main goal is to predict the future performance of an athlete, using its own past performances and information from other athletes. Due to their genuine nature and the uneven observation timestamps, performance data are considered as functional over time and fitted thanks to B-spline basis. These functional observations are assumed to be realizations of a Gaussian stochastic process, and the problem is addressed through Gaussian process regression. This task is treated by a mixed effect model where a mean process is common to all observations and a random process is fitted on each individual curve. This approach allows to rely on information from the whole dataset for modeling and then settles the issue of sparse observations. A first clustering step is performed on the functional data, which allows a cluster-specific prediction step afterwards. The whole procedure is estimated from a Bayesian perspective, which naturally enables the computing of uncertainty and credible intervals for our predictions. The effectiveness of the method is assessed through a simulation study as well as an application on a real dataset from the French Swimming Federation. The usefulness of

[*]MAP5 - Université Paris Descartes | IRMES - INSEP, arthur.leroy@parisdescartes.fr
[†]MAP5 - Université Paris Descartes, servane.gey@parisdescartes.fr
[‡]MAP5 - Université Paris Descartes, pierre.latouche@parisdescartes.fr

the method is twofold, offering a better understanding of the performance progression phenomenon in sport, and providing an automatic talent identification tool.

Keywords. Functional Data, Gaussian Process, Prediction, Elite sport, B-splines, ...

Contexte

Dans le domaine sportif, un point-clé de la continuité d'une discipline au plus haut niveau est la détection des jeunes talents prometteurs. Pour espérer briller dans les futures échéances internationales, les fédérations sportives essayent de développer des outils objectifs pour appuyer les entraineurs dans le processus de sélection. Les données récoltées consistent généralement en un ensemble de performances des athlètes (temps en course) en fonction de leur âge. Les objets étudiés dans ce travail sont appelés *courbes de progression*, où l'on considère les performances recueillies comme des observations ponctuelles d'un processus continu. Chaque athlète possède une courbe de progression individuelle retraçant sa carrière, et ce sont ces fonctions qui seront étudiées pour comparer les sportifs, les regrouper selon leur profil, et tenter de prédire leurs performances futures. Les outils d'analyse de données fonctionnelles, de clustering, et d'apprentissage supervisé ont pour but applicatif de fournir une prédiction fiable, et d'incertitude mesurée, de la courbe de progression à venir d'un jeune athlète.

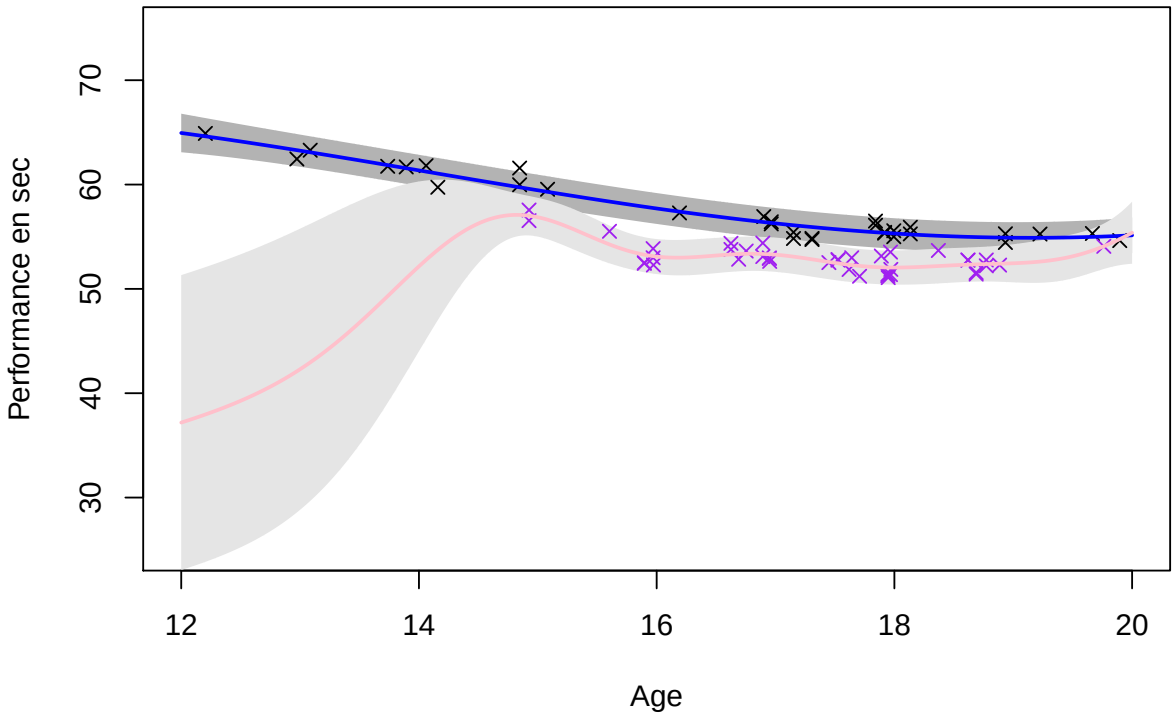
D'un point de vue théorique, les données sportives apportent un lot de difficultés qui sort souvent des cadres classiques des données longitudinales, et forcent la création de modèles spécifiques à cette problématique. En effet, deux caractéristiques principales orientent les choix de modélisation dans cette étude :

- Une seule variable est observée (la performance au cours du temps), avec peu d'occurences par individus (quelques dizaines de points).
- Le nombre d'observations est différent d'un individu à l'autre et les instants d'observations sont différents.

Ces contraintes sont la raison principale du choix de considérer les observations comme fonctionnelles. Par ailleurs, la nécessité de quantifier l'incertitude des prédictions, notamment pour être utilisable en pratique, nous pousse à adopter une approche probabiliste, via la modélisation par processus Gaussien, et le formalisme Bayésien pour l'estimation du modèle.

Modélisation

La modélisation et l'étude de données fonctionnelles est un sujet foisonnant dans la littérature des deux dernières décennies. Deux ouvrages en particulier sont considérés comme des références incontournables, Ramsay and Silverman (2005) avec une approche paramétrique qui étend de nombreuses méthodes multidimensionnelles au cadre fonctionnel, et Ferraty and Vieu (2006) qui prend un point de vu non paramétrique, se concentrant sur des méthodes à noyaux et la définition de métriques pertinentes dans le cadre fonctionnel pour pouvoir appliquer des heuristiques classiques. De part la nature de nos observations très irrégulières, des outils de modélisation présentés dans ces livres, tels que la décomposition dans une base de B-splines, seront utilisés. Cependant, le cadre fréquentiste de ces ouvrages ne permet pas de prendre en compte les incertitudes de modélisation, qui sont importantes compte tenu de notre faible nombre d'observations par individu (quelques dizaines). La modélisation de données fonctionnelles par processus Gaussien est une approche récente qui permet de prendre en compte cette incertitude, comme proposé dans le livre Carl Edward Rasmussen and Williams (2006) ou les articles J. Q. Shi and Wang (2008), Yang et al. (2017). La figure ci-dessous obtenue à l'aide du package R GPFDA (Jian Qing Shi, Choi, and Choi (2011)), illustre la modélisation par processus Gaussien des courbes de progression de deux nageurs français de 100m nage libre.



Cet exemple met l'accent sur la difficulté d'un lissage cohérent des observations, notamment à travers la courbe rose pour laquelle les seules observations disponibles arrivent après 14 ans. Même si l'on observe bien une incertitude dépendante de la quantité de points d'observation, l'approche consistant à définir une fonction pour un individu seulement à partir de ses propres observations est très insatisfaisante. Pour remédier à ce problème, les données fonctionnelles vont être obtenues en utilisant un modèle de régression mixte par processus Gaussiens. Soit $Y_i(t)$ la courbe de progression de l'individu i, on définit:

$$Y_i(t) = \mu(t) + X_i(t) + \epsilon_i$$

Avec:

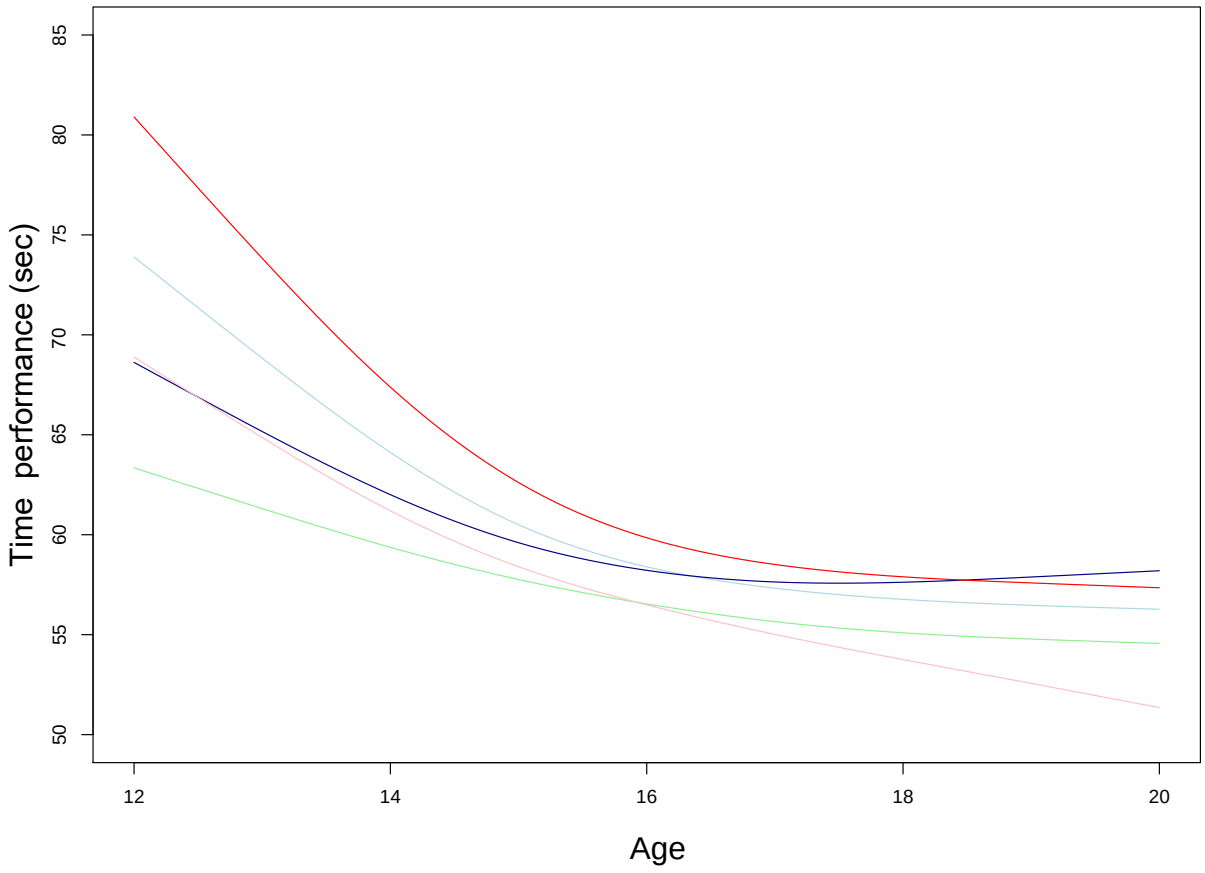
- $\mu(t)$ une fonction moyenne commune à tous les individus,
- $X_i(t) \sim GP(0, \Sigma_i(\cdot, \cdot))$, un processus Gaussien de moyenne nulle et de noyau de covariance $\Sigma_i(\cdot, \cdot)$,
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$, un bruit blanc Gaussien.

Ainsi, y_i est également un processus Gaussien, de moyenne μ, calculée sur tous les individus, et de structure de covariance $\Sigma_i(\cdot, \cdot) + \sigma^2 Id$ estimée via les observations spécifiques à l'individu i. Cette approche permet d'utiliser l'information de tout le jeu de données pour la modélisation individuelle et évite ainsi de définir des courbes abusives pour les athlètes ayant trop peu d'observations sur certains intervalles de temps.

Clustering de courbes

Un premier travail de clustering avait été présenté aux JDS 2018 sur le même jeu de données, et a fait l'objet de l'article Leroy et al. (2018) par la suite. Les nageurs ont ainsi pu être regroupés dans 5 clusters distincts en

fonction de leur niveau de performance et de leur dynamique de progression. Ces résultats (figure ci-dessous) donnent un premier éclairage sur la façon dont les nageurs progressent pendant les premières années de leur carrière et permettent d'utiliser l'information de leur appartenance à un cluster pour effectuer une prédiction plus cohérente.



Travaux en cours

Dans le modèle mixte de régression par processus Gaussien, il est en fait possible de considérer le processus Y_i comme un mélange de processus Gaussiens $Y_i = \sum_{k=1}^{K} \pi_k GP_{ik}(\mu_k, \Sigma_{ik}(\cdot,\cdot))$. Chaque processus ayant une moyenne et un noyau de covariance spécifique, chacun représentant un des K clusters de nos données. Le problème de clustering repose donc sur l'estimation des coefficients π_k, la probabilité d'appartenir à un cluster k, que l'on cherche à estimer également dans notre modèle Bayésien. De telles idées de mélange pour les processus Gaussiens ont été étudié dans Carl E. Rasmussen and Ghahramani (2002) et plus récemment dans Bouveyron, Fauvel, and Girard (2015) pour le clustering.

Régression fonctionnelle par processus Gaussien

Le problème de la régression fonctionnelle par processus Gaussien a été abordé dans une série d'articles de J.Q. Shi, qui propose notamment dans J. Q. Shi and Wang (2008) un modèle pour prédire une sortie fonctionnelle

à partir de covariables fonctionnelles et/ou fini-dimensionnelles. Un modèle hiérarchique est défini et un algorithme EM est utilisé pour estimer les paramètres du modèle, notamment la structure de covariance qui relie les covariables à la fonction à prédire. L'article Yang et al. (2017) propose une généralisation de ce modèle en rajoutant un niveau de hiérarchie dans la définition des paramètres. En effet, le processus moyen, maintenant considéré comme aléatoire, est lui-même un processus Gaussien a priori, et une loi a priori inverse-Wishart (IW) est définie sur les noyaux de covariance pour éviter d'imposer une structure particulière.

Travaux en cours

En s'appuyant sur ces travaux précédents, l'idée est d'utiliser un modèle hiérarchique Bayésien qui va à la fois estimer une structure de covariance globale pour le processus moyen μ et une covariance individuelle pour $\Sigma_i(\cdot, \cdot)$. Les courbes de progression des nageurs étant régulières et globalement monotones, un noyau de covariance exponentiel-polynomial paraît suffisant, sans recourir à la définition d'un a priori IW. La principale différence avec les travaux précédents venant du fait que les courbes de progression sont à la fois les entrées et la sortie du modèle, et les observations étant irrégulières, une nouvelle procédure d'estimation du modèle doit être proposée, à travers la définition d'un algorithme Monte Carlo Markov Chain (MCMC) spécifique. Ce travail est actuellement en cours et sera le principal résultat d'intérêt de la présentation.

Références

Bouveyron, Charles, Mathieu Fauvel, and Stephane Girard. 2015. "Kernel Discriminant Analysis and Clustering with Parsimonious Gaussian Process Models." *Statistics and Computing* 25 (6): 1143–62. doi:10.1007/s11222-014-9505-x.

Ferraty, Frédéric, and Philippe Vieu. 2006. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.

Leroy, Arthur, Andy Marc, Olivier Dupas, Jean Lionel Rey, and Servane Gey. 2018. "Functional Data Analysis in Sport Science: Example of Swimmers' Progression Curves Clustering." *Applied Sciences* 8 (10): 1766. doi:10.3390/app8101766.

Ramsay, James O., and Bernard W. Silverman. 2005. *Functional Data Analysis*. Springer.

Rasmussen, Carl E., and Zoubin Ghahramani. 2002. "Infinite Mixtures of Gaussian Process Experts." In *Advances in Neural Information Processing Systems 14*, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani, 881–88. MIT Press.

Rasmussen, Carl Edward, and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press.

Shi, J. Q., and B. Wang. 2008. "Curve Prediction and Clustering with Mixtures of Gaussian Process Functional Regression Models." *Statistics and Computing* 18 (3): 267–83. doi:10.1007/s11222-008-9055-1.

Shi, Jian Qing, Taeryon Choi, and Taeryon Choi. 2011. *Gaussian Process Regression Analysis for Functional Data*. Chapman and Hall/CRC. doi:10.1201/b11038.

Yang, Jingjing, Dennis D. Cox, Jong Soo Lee, Peng Ren, and Taeryon Choi. 2017. "Efficient Bayesian Hierarchical Functional Data Analysis with Basis Function Approximations Using Gaussian-Wishart Processes: Efficient Bayesian Hierarchical Functional Data Analysis." *Biometrics* 73 (4): 1082–91. doi:10.1111/biom.12705.