ESTIMATION NON-PARAMÉTRIQUE D'UNE RÉGRESSION SPHÉRIQUE DANS LE CADRE DE L'ANALYSE DE DONNÉES FONCTIONNELLES

Papa Alioune Meissa MBAYE [1,2] & Chafik SAMIR [1] & Anne-Françoise YAO [2]

[1] *Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, Université Clermont Auvergne*

[2] *Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne*

papa_alioune_meissa.mbaye@uca.fr[1,2]*, chafik.samir@uca.fr*[1]*, anne.yao@uca.fr*[2]

Résumé. L'analyse de données à valeurs dans une sous-variété riemannienne $\mathcal{M}$, de dimension finie k suscite beaucoup d'intérêt dans plusieurs domaines de la science et plus particulièrement pour analyser des données médicales. Ces données peuvent être des courbes, des contours ou des volumes. Dans ce travail, nous nous intéressons à la régression où le prédicteur est à valeurs dans un espace de fonctions. Nous nous intéressons en particulier au cas où cet espace est non linéaire, plus précisément la sphère unité $\mathcal{S}^k$. L'estimateur utilisé est un estimateur à noyau. Ce dernier est appliqué à différentes problématiques. Les données (courbes) étudiées peuvent être sujettes au problème de recalage.
Au cours de notre exposé, nous aborderons ce problème dans un premier temps puis celui du choix d'une représentation fonctionnelle optimale. Enfin, nous illustrerons notre méthode par des applications à des données simulées ainsi qu'à des données réelles.

Mots-clés. Analyse de données fonctionnelles, Régression, Méthodes à noyaux, Recalage, Sphère.

Abstract. Analysis of data belonging to a k-dimensional Riemannian manifold $\mathcal{M}$ raises interests in different scientific areas, particularly in medicine. Data can be curves, contours or volumes. In this paper, we focus on regression on what predictors are in a function space. We emphasize on the case in which this space is nonlinear, more precisely the sphere $\mathcal{S}^k$. The estimateur used is the kernel estimator. This latter is applied to different problems. Data (curves) studied can be subject to registration problem.
During our presentation, we will address this issue at first, and at second to choose a optimal functional representation. Finally, we will illustrate our words by applications on simulated and real data.

Keywords. Functional Data Analysis, Regression, Kernel methods, Registration, Sphere.

1 Introduction

Faire l'analyse statistique sur des variétés de dimension k, $k \in \mathbb{N}^*$, est un sujet qui est en pleine expansion qui requiert souvent quelques prétraitements de données, comme par exemple le

recalage de courbes ou d'images. Il est important de noter aussi que les variétés ont une structure plus complexe que celle de l'espace euclidien $\mathbb{R}^n$. Dans ce travail, nous nous intéressons à l'estimation non paramétrique de la fonction de régression. Nous commençons par présenter la méthode de recalage utilisée. Ensuite, nous détaillerons notre approche. Enfin, nous présentons quelques résultats obtenus.

2 Formulation du problème

Soit une suite d'observations (f_i, y_i), $i = 1, ..., n$, du couple de variables aléatoires (f, Y) où f représente la variable explicative et Y la variable à expliquer. f est une variable aléatoire fonctionnelle à valeurs dans un espace métrique $(\mathcal{H}, d)$ et Y est une variable binaire ($Y \in \{0, 1\}$). Notre objectif est d'estimer la fonction de lien définie par:

$$Y = r(f) + \epsilon, \ \ \forall f \in \mathcal{H} \tag{1}$$

avec r la fonction de régression et $\epsilon \sim \mathcal{N}\ (0, \sigma^2)$. Comme annoncé précédemment, nous commençons par un prétraitement des données, le recalage.

2.1 Introduction au recalage

Le décalage temporel dans les données peut être causé par plusieurs facteurs. Si on considère par exemple les courbes issues de la force de la main (voir Figure 1), nous constatons que pour une même personne, on peut avoir deux courbes différentes. Il existe beaucoup de travaux sur les données fonctionnelles dans la littérature mais peu tiennent compte des décalages temporels qui peuvent apparaître sur les données, alors que le recalage est un prétraitement fondamental pour réduire les erreurs d'estimations. Plusieurs méthodes de recalage existent, mais l'objectif reste le même : supprimer la variabilité temporelle. Un exemple d'algorithme de recalage peut se résumer comme suit : soient les deux ensembles de fonctions $\{f_i, i = 1, ..., n\}$ et $\{\gamma_i, i = 1, ..., n\}$ telles que $\gamma : [0, 1] \to [0, 1], \gamma(0) = 0, \gamma(1) = 1$ et sa dérivée $\gamma' \succ 0$. L'objectif est de trouver une suite de fonctions $\{\gamma_i^*, i = 1, \ldots, n\}$ telles que $\{f_i \circ \gamma_i^*, i = 1, \ldots, n\}$ soient alignées et ne varient qu'en amplitude.
Une illustration de cet aspect est donnée à la Figure 1 qui représente les courbes issues de la force de la main avant et après recalage.

2.2 Régression par méthode à noyau

Notre objectif est d'estimer la fonction de régression r de l'équation (1) par un estimateur à noyau définie par :

$$\hat{r}(f) = \frac{\sum_{i=1}^{n} y_i K_h(d(f, f_i))}{\sum_{i=1}^{n} K_h(d(f, f_i))}, \ \ \forall f \in \mathcal{H} \tag{2}$$

où n est le nombre d'observations et $K_h(d(f,f_i)) = K\left(\frac{d(f,f_i)}{h}\right)$. K est un noyau, h est la fenêtre (qui tend vers 0 lorsque n tend vers l'infini) et d est une métrique sur une sous-variété de $\mathcal{H}$.
Dans ce qui suit, cette sous-variété est soit un espace linéaire qu'on note $\mathcal{H}_1 = \{f \in C([0,1]), \|f\|^2_{\mathcal{H}_1} < \infty\}$ ou non linéaire $\mathcal{H}_2 = \{f \in \mathbb{L}^2([0,1], <, >), \|f\|_{\mathcal{H}_1} = 1\}$.
Dans chaque cas, nous nous intéressons à la performance de l'estimateur pour les 3 représentations suivantes : la courbe initiale f, la vélocité f' et l'accélération f''. f' et f'' nous permettent respectivement d'avoir les variations au niveau des courbes initiales et celles au niveau de la vélocité.

3 Applications à la classification des courbes

Nous illustrons notre propos à travers des données simulées d'une part et d'autre part sur des données réelles (Figure 1). Les méthodes à noyaux nécessitent l'optimisation de plusieurs paramètres : la métrique d et la fenêtre h. Le noyau K utilisé est le noyau gaussien. Le choix de la métrique dépend de l'espace $\mathcal{H}$ dans lequel sont définies les f_i. Si $\mathcal{H} = \mathcal{H}_1$, d représente la métrique $\mathbb{L}^2$ et si $\mathcal{H} = \mathcal{H}_2$, d est une géodésique. En ce qui concerne le choix de la fenêtre optimale, nous choisirons la valeur de h qui maximise le critère MCC (Matthews Correlation Coefficient). Il prend des valeurs comprises entre -1 et 1 et une classification est bonne si la valeur de MCC est proche de 1. Un extrait des résultats concernant les données réelles est donné dans les tableaux 1 et 2.

Input / Metric	Initial	Velocity	Acceleration
$\mathbb{L}^2$	30.24%	24.10%	24.5%
$\mathbb{L}^2$/recalage	30.8%	28.9%	17.5%

Table 1: Taux d'erreurs obtenus sur $\mathcal{H}_1$.

Input / Metric	Initial	Velocity	Acceleration
Geodesique	25.54%	19.9%	14.63%
Geodesique/recalage	30.31%	12.48%	11.45%

Table 2: Taux d'erreurs obtenus sur $\mathcal{H}_2$.

Ce dernier montre que les résultats sont meilleurs quand les f_i appartiennent à l'espace non linéaire $\mathcal{H}_2$ et que le recalage a nettement amélioré les résultats de la classification sur $\mathcal{H}_2$ pour les représentations vélocité et accélération.

3.1 Graphiques

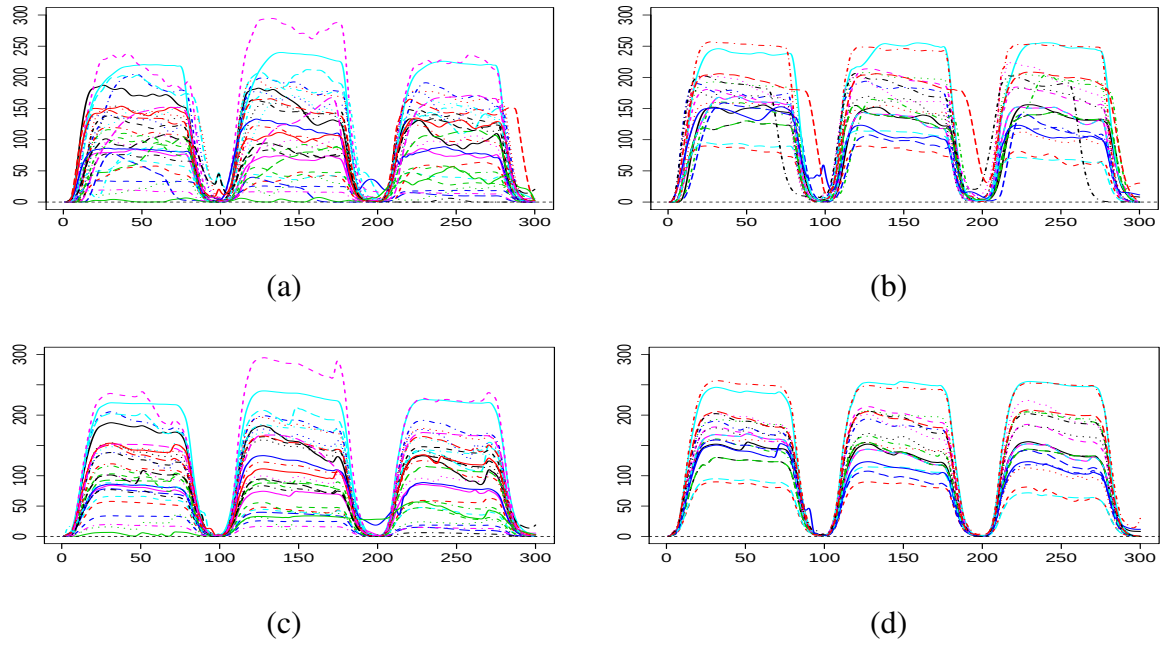

Figure 1: **(a)** Courbes de 30 personnes malades avant recalage, **(b)** Courbes de 20 personnes saines avant recalage, **(c)** Courbes de 30 personnes malades après recalage et **(d)** Courbes de 20 personnes saines après recalage.

Quelques références

Ramsay, J. O. and Silverman, B. W. (2005), Functional Data Analysis, *Second Edition, Springer Series in Statistics.*

Ferraty, F. and Vieu, P. (2006), Nonparametric functional data analysis: theory and practice, *Springer Science and Business Media.*

Pelletier, B. (2006), Non-parametric regression estimation on closed Riemannian manifolds, *Journal of Nonparametric Statistics 18.1 : pp. 57-67.*

Tucker, J. D. (2014), Functional component analysis and regression using elastic methods, *Electronic Theses, Treatises and Dissertations, Florida State University*.

Samir, C. Kurtek, S. Srivastava, A. and Borges, N. (2016), An elastic functional data analysis framework for preoperative evaluation of patients with rheumatoid arthritis, *IEEE WACV*.

Srivastava, A. and Klassen, E.P. (2016), Functional and shape data analysis, *New York: Springer*.

Lin, Z. and Yao, F. (2018), Intrinsic Riemannian Functional Data Analysis, *arXiv preprint arXiv:1812.01831*.

Cet article est réalisé dans le cadre d'un projet financé par la région Auvergne-Rhône-Alpes.