Alignement de courbes avec un modèle mixte fonctionnel

Emilie Devijver [1] & Gerda Claeskens [1] & Irène Gijbels [1]

[1] *Department of Mathematics and Leuven Statistics Research Center, Katholieke Univsersiteit Leuven, Belgium, emilie.devijver@kuleuven.be*

Résumé. Les données fonctionnelles présentent différentes variabilités qu'il faut prendre en compte dans la modélisation. Dans ce projet, nous travaillons sur des techniques d'alignement de courbes, en prenant en compte la variabilité individuelle. On propose pour se faire de travailler avec un modèle mixte fonctionnel non linéaire. Chaque observation d'un processus fonctionnel est alors décomposée comme la somme d'une courbe moyenne et une courbe individuelle, le tout altéré par composition avec une fonction d'alignement. La fonction d'alignement est supposée paramétrique, et on estime les paramètres avec un modèle mixte linéaire. Chaque fonction du modèle est décomposée sur une base de splines. Ce modèle général permet de capter différents types de variabilité, et de faciliter l'interprétation. Dans cet exposé, je proposerai une méthode d'estimation des paramètres de ce modèle, et j'illustrerai la méthode sur des données simulées et des données benchmark.

Mots-clés. Données fonctionnelles, alignement, modèle à effets mixtes.

Abstract. Functional data usually present two types of variability : a phase variability and an amplitude variability. In this work, we propose to use a non linear fonctional mixed model to carry both variabilities, in order to improve the modeling and the interpretation. Each observation is then supposed to be the sum of a mean curve and an individual curve, both composed with a parametric warping function. Every curve is decomposed onto a B-spline basis. In this talk, I will describe a procedure to estimate the parameters, and I will illustrate the method on simulated and benchmark data.

Keywords. Functional data, warping, mixed effect model.

1 Le modèle

On considère le modèle mixte fonctionnel suivant : pour un individu $i \in \{1, \dots, n\}$, pour un instant $t \in [0, 1]$,

$$Y_i(t) = \mu \left\{ w^{-1}(t; \boldsymbol{\theta}_i) \right\} + U_i \left\{ w^{-1}(t; \boldsymbol{\theta}_i) \right\} + \boldsymbol{\varepsilon}_i \left\{ w^{-1}(t; \boldsymbol{\theta}_i) \right\} ;$$

où μ représente la moyenne commune, U_i l'effet aléatoire associé à l'individu i, et $w : [0, 1] \to [0, 1]$ la fonction d'alignement, strictement croissante, et dépendant d'un paramètre $\boldsymbol{\theta}_i$ inconnu.

On suppose qu'on observe une discrétisation des fonctions $(Y_i)_{1 \le i \le n}$ sur les points fixes $(t_{i,j})_{1 \le j \le T_i, 1 \le i \le n}$: $(y_{i,j})_{1 \le j \le T_i, 1 \le i \le n} = (Y_i(t_{i,j}))_{1 \le j \le T_i, 1 \le i \le n}$. On considère alors le modèle suivant :

$$Y_i(t_{i,j}) = \mu \left\{ w^{-1}(t_{i,j}; \boldsymbol{\theta}_i) \right\} + U_i \left\{ w^{-1}(t_{i,j}; \boldsymbol{\theta}_i) \right\} + \varepsilon_{i,j}. \tag{1}$$

On décompose chaque signal sur une base de splines, et on note $\boldsymbol{\alpha}^{\mu}$ et $\boldsymbol{\alpha}^{U_i}$ les coefficients respectifs de μ et U_i.

La fonction d'alignement considérée est la suivante : pour tout $t \in [0, 1]$,

$$w^{-1}(t; \boldsymbol{\theta}_i) = \frac{\int_0^t \exp \left\{ h^{-1}(u; \boldsymbol{\theta}_i) \right\} du}{\int_0^1 \exp \left\{ h^{-1}(u; \boldsymbol{\theta}_i) \right\} du}, \qquad \text{où} \qquad h^{-1}(u; \boldsymbol{\theta}_i) \ = \sum_{l=-p_h}^{K_h} \theta_{i,l} \bar{B}^h_{l,p_h+1}(u; \boldsymbol{\kappa}^h).$$

avec $(\bar{B}^h_{l,p_h+1})_{-p_h \le l \le K_h}$ une base de splines où chaque fonction est centrée, i.e. $\int_0^1 \bar{B}^h_{l,p_h+1}(u; \boldsymbol{\kappa}^h) du = 0$.

On suppose que les paramètres $\boldsymbol{\theta}_i$ sont aléatoires, tels que

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \mathbf{E}_i + \tilde{\boldsymbol{\varepsilon}}_i$$

avec $\mathbf{E}_i \sim \mathcal{N}_r(\mathbf{0}_r, \Sigma^{\mathbf{E}})$ et $\tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}_r(\mathbf{0}_r, \sigma^2_{\tilde{\boldsymbol{\varepsilon}}} \mathbf{I}_r)$. Pour éviter des problèmes d'identifiabilité, on suppose que la variance du bruit $\sigma^2_{\tilde{\boldsymbol{\varepsilon}}}$ est connue. On notera $\Sigma^{\boldsymbol{\theta}}$ la variance associée à $(\boldsymbol{\theta}_i)_{1 \le i \le n}$, $\Sigma^{\boldsymbol{\theta}} = \Sigma^{\mathbf{E}} + \sigma^2_{\tilde{\boldsymbol{\varepsilon}}} \mathbf{I}_r$.

2 La procédure d'estimation

On suppose donc qu'on observe n courbes discrétisées $(y_{i,j})_{1 \le j \le T_i, 1 \le i \le n}$, la courbe i étant évaluée sur T_i points.

On fixe les noeuds associés à la décomposition sur une base de splines.

1. Initialisation des paramètres d'alignement :

(a) calcul de la *deepest function* $\hat{\mu}^{(0)} = \mu^{\text{deep}}$. On ne considère pas les effets aléatoires U_i dans l'initialisation ;

(b) pour tout $i \in \{1, \ldots, n\}$, on approche les paramètres $\boldsymbol{\theta}_i$ en minimisant un critère L_2 empirique entre $y_i \circ w_{\hat{\boldsymbol{\theta}}_i^{(0)}}$ et μ^{deep} ;

(c) on considère un modèle à effets mixtes linéaire sur ces observations approchées

$$\hat{\boldsymbol{\theta}}_i^{(0)} = \hat{\boldsymbol{\theta}}_0^{(0)} + \hat{\mathbf{E}}_i^{(0)} + \tilde{\boldsymbol{\varepsilon}}_i;$$

et on en déduit $\hat{\boldsymbol{\theta}}_0^{(0)}$, $(\hat{\Sigma}^{\boldsymbol{\theta}})^{(0)}$ et $\hat{\boldsymbol{\theta}}_i^{(0)} = \hat{\boldsymbol{\theta}}_0^{(0)} + \hat{\mathbf{E}}_i^{(0)}$.

2. Itérations jusqu'à convergence : pour l'étape (ite) $\in \{1, \ldots, \text{end}\}$,

(a) sachant $(\hat{\boldsymbol{\theta}}_i^{(\text{ite}-1)}, \hat{\boldsymbol{\theta}}_0^{(\text{ite}-1)}, (\hat{\Sigma}^{\boldsymbol{\theta}})^{(\text{ite}-1)})$:

- aligner les courbes observées avec $w^{-1}_{\hat{\boldsymbol{\theta}}_i^{(\text{ite}-1)}}$;
- estimer $\{(\hat{\boldsymbol{\alpha}}^{\mu})^{(\text{ite})}, (\hat{\boldsymbol{\alpha}}_i^{U})^{(\text{ite})}, (\hat{\Sigma}^{U_i})^{(\text{ite})}, (\hat{\sigma}_{\boldsymbol{\varepsilon}}^2)^{(\text{ite})}\}$,

(b) sachant $(\hat{\mu}^{(\text{ite})}, \hat{U}_i^{(\text{ite})}, (\hat{\Sigma}^{U_i})^{(\text{ite})}, (\hat{\sigma}_{\boldsymbol{\varepsilon}}^2)^{(\text{ite})})$:

- pour tout $i \in \{1, \ldots, n\}$, approcher $\boldsymbol{\theta}_i$ par $\boldsymbol{\theta}_i^{(\text{ite})}$ qui minimise le critère des moindres carrés pondéré entre y_i et la moyenne courante $\mu^{(\text{ite})} + U_i^{(\text{ite})}$;
- estimer un modèle à effets mixtes linéaire sur ces observations :

$$\boldsymbol{\theta}_i^{(\text{ite})} = \hat{\boldsymbol{\theta}}_0^{(\text{ite})} + \hat{\mathbf{E}}_i^{(\text{ite})} + \tilde{\boldsymbol{\varepsilon}}_i;$$

- définir $\hat{\boldsymbol{\theta}}_i^{(\text{ite})} = \hat{\boldsymbol{\theta}}_0^{(\text{ite})} + \hat{\mathbf{E}}_i^{(\text{ite})}$.

3 Données simulées et données réelles

On comparera notre méthode à celle présentée dans [1] sur des donnée simulées.

On présentera aussi les résultats sur les données Pinch disponibles dans le package R `fda` présentées dans la Figure 1.

Dans la Figure 2, on représente les courbes alignées pour montrer la variabilité en amplitude, et dans la Figure 3 on représente les fonctions d'alignement pour montrer la variabilité en amplitude.

Bibliographie

[1] D. Gervini et P.A. Carter, *Warped functional analysis of variance*, Biometrics, 70(3):526-535, 2014.

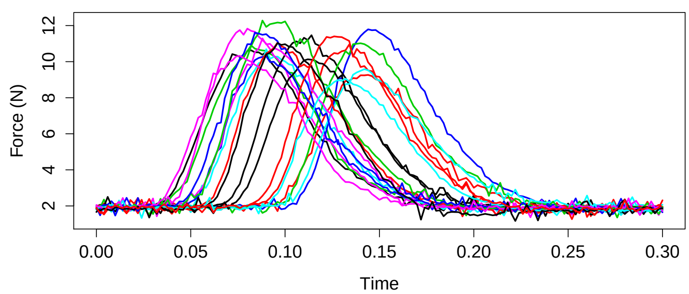

Figure 1: Données Pinch

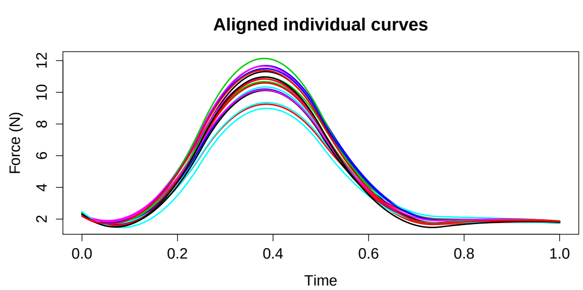

Figure 2: Données alignées

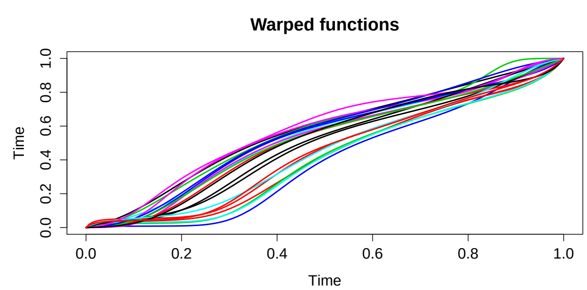

Figure 3: Fonction d'alignement estimées