

ESTIMATION DANS LES MODÈLES MIXTES FONCTIONNELS EN PRÉSENCE DE DÉFORMATIONS INDIVIDUELLES NON-LINÉAIRES

Gerda Claeskens¹, Madison Giacomini^{2,*}, Irène Gijbels² and Maarten Jansen³

¹ *ORSTAT & Leuven Statistics Research Center, KU Leuven;
gerda.claeskens@kuleuven.be*

² *Department of Mathematics & Leuven Statistics Research Center, KU Leuven;
joycemadison.giacomini@wis.kuleuven.be, irene.gijbels@wis.kuleuven.be*

³ *Departments of Mathematics and Computer Science, ULB, maarten.jansen@ulb.ac.be
* Corresponding author*

Résumé. Nous nous intéressons à l'étude de données fonctionnelles, mesurées de manière répétée sur un ensemble d'individus. Dans un contexte multi-individus, les courbes individuelles sont souvent mesurées sur une discrétisation propre, non régulière, et il est courant d'observer des variations inter-individuelles aussi bien en amplitude (*i.e.* taille des principales caractéristiques des courbes) qu'en phase (*i.e.* timing des ces caractéristiques). Il existe une vaste littérature concernant l'étude séparée de chacun de ces types de variations, cependant, leur prise en compte simultanée reste un sujet relativement peu abordé.

Nous proposons une nouvelle procédure basée sur une modélisation fonctionnelle mixte des signaux, intégrant des effets aléatoires fonctionnels modélisant les variations en amplitude autour d'un effet fixe fonctionnel moyen, et des fonctions de déformations (warping) aléatoires individuelles modélisant les variations inter-individuelles en phase. Notre approche consiste alors en une décomposition des effets fixes et aléatoires dans une base d'ondelettes, permettant la considération de signaux spatialement inhomogènes, tandis que les déformations individuelles sont interpolées à l'aide de splines cubiques monotones d'Hermite, offrant une relation directe entre leurs noeuds et les caractéristiques des courbes étudiées.

L'estimation des paramètres du modèle est alors réalisée par maximum de vraisemblance, à l'aide de l'algorithme MCEM, une variante de l'algorithme EM, consistant à remplacer l'étape E, dont le calcul est rendu difficile par la non-linéarité des fonctions de déformations, par une approximation de Monte-Carlo.

Le comportement général de notre procédure est étudié sur une vaste étude de simulation et nous illustrons notre méthode sur un jeu de données réelles.

Mots-clés. Ondelettes, Déformations non-linéaires, Modèles mixtes non-linéaires, Algorithme MCEM.

Abstract. We consider data that consist of set of curves recorded on individuals, modelled as functional data. In a multi-individual context, curves are often recorded on subject-specific non-regular time grids and variations in both phase (i.e variations in timing of features) and in amplitude (i.e variations in size of features) are commonly encountered. While individual sample effects and curve alignment are widely studied separately in the functional setting, only little attention has been devoted to the study of simultaneous amplitude and phase variations.

We propose a functional warped mixed model that accounts for individual deviations in amplitude around a main functional pattern by considering functional random effects whereas modelling of individual phase deviations is done through nonlinear warping functions. We then use a wavelet representation of both the main pattern and individual amplitude variations, allowing to deal with spatially heterogeneous curves. Non-linear subject-specific phase variations are modelled through a monotone Hermite spline decomposition which offer a close relation with the landmark structure.

In this setting, we propose a new maximum-likelihood based estimation method using a MCEM algorithm, a variant of the EM algorithm. The non-linearity of the mixed model yields the E-step untractable and a Monte-Carlo integration is used for random effects predictions of individual amplitude deviation coefficients and individual warping parameters.

The performance of the procedure is investigated through an extensive simulation study and we illustrate our method on real data examples.

Keywords. Wavelets, Nonlinear Warping functions, Nonlinear Mixed Effects Models, EM algorithm

1 Introduction

De nos jours, les données fonctionnelles sont fréquemment rencontrées dans de nombreux domaines d'applications, par exemple en biologie moléculaire, en économie ou en linguistique. La modélisation de telles données offre de nouveaux challenges car celle-ci doit être suffisamment flexible tout en restant estimable. Dans ce cadre, les ondelettes ont grandement gagné en popularité au cours des dernières décennies, notamment grâce à leur capacité à modéliser des courbes spatialement inhomogènes. Lorsque l'on considère une courbe mesurée de manière régulière sur une grille dyadique $(t_j)_{j=1,\dots,T}$ (*i.e.*, $t_j = j/2^J$, pour J entier), de nombreuses techniques numériquement efficaces ont été développées, parmi lesquelles, la transformée en ondelettes discrètes de Mallat ou le concept de seuillage introduit par Donoho et Johnstone. Ces méthodes tirent parti de l'orthonormalité des bases considérées et de leur structure en arbre mais requièrent des contraintes dyadiques sur la grille de design considérée. Cependant, les données issues d'applications réelles remplissent rarement ce type de conditions. De plus, en présence de plusieurs individus, il est courant d'être confronté à des variations inter-individuelles, d'une part, vis-à-vis de l'amplitude des caractéristiques principales des signaux considérés mais aussi, d'autre part, vis-à-vis du timing de ces caractéristiques entre les différents individus.

La prise en compte des variations en amplitude entre les individus est en général réalisée à travers l'utilisation de modèles mixtes fonctionnels. Dans ce cadre, au sein d'un groupe homogène d'individus, les signaux individuels sont décomposés en un effet fixe fonctionnel principal modélisant le comportement moyen au sein du groupe tandis que l'introduction d'effets aléatoires fonctionnels individuels modélisent la présence de déviations fonctionnelles autour de l'effet fixe principal, spécifiques à l'individu. Dans une approche basée sur les ondelettes, les travaux de Morris et Carroll (2006) ou Antoniadis et Sapatinas (2007) permettent de considérer à l'aide d'ondelettes des effets fixes et aléatoires situés dans le même espace fonctionnel, offrant ainsi une meilleure interprétabilité.

Parallèlement, l'étude des problématiques d'alignement de courbes par rapport à leurs principales caractéristiques est aussi un sujet largement étudié depuis quelques années par l'intermédiaire de fonctions de déformations (ou fonctions de warping). Les difficultés de ces modèles résident principalement dans les contraintes de modélisation des fonctions de déformations : en effet, de telles transformations doivent être inversibles, monotones, tout en restant identifiables et en s'adaptant aux particularités des courbes étudiées. Ainsi, Gervini et Gasser (2004) proposent une modélisation des ces déformations basée sur les splines, Claeskens et al. (2010) une approche basée sur des fonctions de warping localisées et multi-échelles ou encore Bigot (2013), offrant une modélisation des fonctions de déformations comme solutions d'équations différentielles ordinaires, pour n'en citer que quelques-unes. Cependant, le point commun de ces approches est de proposer ensuite des procédures d'estimation des paramètres basées sur des critères de type moindres carrés, les rendant peu adaptées au cadre des modèles mixtes fonctionnels.

Jusqu'à présent, la pratique courante lors de l'étude des données présentant des

déviations inter-individuelles en phase et en amplitude, est d'appliquer en premier lieu une procédure d'alignement afin de synchroniser les données individuelles. L'étape d'alignement est alors considérée comme une étape de pré-traitement, préalable à l'analyse finale, négligeant alors l'incertitude liée à cette dernière. Dans la perspective de traiter simultanément ces deux types de variations, Bigot et Charlier (2011) adoptent une approche basée sur la notion de moyenne de Fréchet, tandis que Gervini et Carter (2014) proposent une approche de type maximum de vraisemblance, basée sur une décomposition de Karhunen-Loève des signaux. Ces méthodes présentent néanmoins certaines limites lorsque les signaux considérés présentent de fortes déviations en amplitude ou présentent des irrégularités.

Dans cet article, nous proposons une nouvelle procédure permettant l'estimation dans un modèle fonctionnel permettant la prise en compte simultanée de variations inter-individuelles en phase et en amplitude.

2 Modèle adopté

Les données considérées sont constituées de n courbes individuelles, notées $y_i(t)$ pour $i = 1, \dots, n$, mesurées sur des grilles de discrétisation non-nécessairement régulières et propre aux individus $(t_{im})_{m=1, \dots, T_i}$, incluses dans l'intervalle $[0, 1]$. Afin de prendre en compte les déviations individuelles en phase et en amplitude, nous adoptons le modèle fonctionnel suivant :

$$y_i(t_{im}) = f_i(w_i(t_{im})) + E_{ij}, \quad (1)$$

où f_i représente le signal sous-jacent associé à l'individu i tandis que la fonction $w_i : [0, 1] \rightarrow [0, 1]$ est un difféomorphisme strictement croissant, modélisant la déformation temporelle subie par le i -ième signal, et traditionnellement désignée comme fonction de warping. Cette dernière est supposée être paramétrée par un vecteur d'effets aléatoires θ_i , tel que $w_i(\cdot) = w(\cdot, \theta_i)$. Les variables aléatoires $(E_{ij})_{i,j}$ sont supposées être indépendantes, centrées et gaussiennes de variance σ^2 . Partant de la modélisation (1), au sein d'un groupe homogène d'individus, les déviations individuelles autour du comportement fonctionnel moyen sont modélisées en décomposant les signaux individuels f_i , $i = 1, \dots, n$, de la manière suivante :

$$f_i(t) = \mu(t) + U_i(t), \quad (2)$$

où μ représente l'effet fixe fonctionnel principal au sein du groupe d'individus tandis que les fonctions U_i représentent les déviations individuelles en amplitude autour de l'effet fixe principal. Les fonctions U_i sont modélisés comme étant des réalisations d'un processus gaussien centré de covariance $K(t, t')$.

Afin couvrir un large champ de régularités fonctionnelles, nous adoptons une approche non paramétrique basée sur les ondelettes pour la représentation des effets fixes et aléatoires. Les coefficients associés à la décomposition des effets aléatoires $\mathbf{b}_i$ sont supposés

gaussiens, centrés et de matrice de covariance $\mathbf{\Gamma}$. A l’instar de Antoniadis et Sapatinas (2007), nous adoptons une modélisation de la matrice $\mathbf{\Gamma}$ permettant d’assurer que les effets fixes et aléatoires se situent dans le même espace fonctionnel et donc, partage la même régularité : ainsi, $\mathbf{\Gamma}$ est supposée diagonale et ses termes de variances sont supposés avoir une décroissance exponentielle vis-à-vis du niveau de résolution.

Les fonctions de warping individuelles sont quant à elles modélisées à l’aide de splines cubiques monotones d’Hermite, dont le principal avantage réside dans le lien direct existant entre leurs noeuds et les principales caractéristiques des fonctions sous-jacentes. Les paramètres aléatoires associés aux fonctions de warping $(\theta_i)_{i=1,\dots,n}$, devant répondre à une contrainte de monotonie, sont alors transformés à l’aide d’une transformée inversible afin de se ramener à une modélisation gaussienne additive non-contrainte.

3 Estimation par maximum de vraisemblance

Au sein des modèles mixtes, l’algorithme EM représente un outil classique d’estimation par maximum de vraisemblance. Dans notre contexte, deux types d’effets aléatoires (ou variables latentes) interviennent : les effets liés aux variations individuelles en amplitude et ceux liés aux variations individuelles en phase. En raison de la non-linéarité des fonctions de déformations, le modèle que nous considérons se place dans la classe des modèles non-linéaires à effets mixtes et de ce fait, l’espérance conditionnelle de la vraisemblance, correspondant à l’étape E de l’algorithme, ne possède pas d’expression explicite. Afin de remédier à cela, suivant Walker (1996), nous utilisons l’algorithme MCEM, une variante de l’algorithme EM, consistant à remplacer le calcul de l’espérance au sein de l’étape E par une approximation de Monte-Carlo.

4 Résultats sur données simulées et données réelles

Afin de comparer les performances de notre méthode à celle des méthodes existantes, nous proposons une procédure permettant de créer des jeux de données synthétiques de manière automatique. Sur cette base, nous comparons alors les méthodes considérées pour différents niveaux de variabilités en phase et en amplitude et pour différents niveaux de bruit. L’intérêt de notre procédure est aussi souligné par une application à des données de spectrométrie de masse.

Bibliographie

References

- [1] Antoniadis, A. and Sapatinas, T. (2007), Estimation and inference in functional mixed-effects models, *CSDA*, 51 (10), p4793-4813.
- [2] Bigot, J. (2013), Fréchet means of curves for signal averaging and application to ECG data analysis, *Annals of Applied Statistics*, 7, 2384–2401.
- [3] Bigot, J. and Charlier, B. (2011), On the consistency of Fréchet means in deformable models for curve and image analysis, *Electronic Journal of Statistics*, 5, 1054–1089.
- [4] Claeskens, G. and Silverman, B. and Slaets, L. (2010), Time Warping achieved by a Bayesian prior-posterior transfer fitting strategy, *Journal of the Royal Statistical Society, Ser. B*, 72 (5), p673–694.
- [5] Gervini, D. and Carter, P.A. (2014), Warped functional analysis of variance, *Biometrics*, 70(3), 526–535.
- [6] Gervini, D. and Gasser, T. (2004), Self-Modeling warping functions, *Journal of the Royal Statistical Society, Ser. B*, 66, 959–971.
- [7] Morris, J.S. and Carroll, R.J. (2006), Wavelet-based functional mixed models, *Journal of the Royal Statistical Society Series B Stat Methodol*, 68, 179–199.
- [8] Walker, S. (1996), An EM algorithm for Nonlinear Random Effects Models, *Biometrics*, 52(3), 934–944