Nonlinear mixed effects modeling and warping for functional data using B-splines

Gerda Claeskens, Emilie Devijver, Irène Gijbels

Nonlinear mixed effects modeling and warping for functional data using B-splines

Gerda Claeskens[a], **Emilie Devijver**[b] and **Irène Gijbels**[c]
gerda.claeskens@kuleuven.be; emilie.devijver@univ-grenoble-alpes.fr; irene.gijbels@kuleuven.be[1]

[a] ORStat and Leuven Statistics Research Center, KU Leuven, Belgium
[b] CNRS, Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, France
[c] Department of Mathematics, Leuven Statistics Research Center, KU Leuven, Belgium

Abstract

In functional data the interest is to find a global mean pattern, but also to capture the individual curve differences in phase and amplitude. This can be done conveniently by building in random effects on two levels: in the warping functions to account for individual phase variations; and in the linear structure to deal with individual amplitude variations. Via an appropriate choice of the warping function and B-spline approximations, estimation in the nonlinear mixed effects functional model is feasible, and does not require any prior knowledge on landmarks for the functional data. Sufficient and necessary conditions for identifiability of the flexible model are provided. A theoretical study is conducted: we establish asymptotic normality and consistency of the estimators of the registration and amplitude models, convergence of the iterative process, and consistency of the final estimator provided by the iterative process. The finite-sample performance of the proposed estimation procedure is investigated in a simulation study, which includes comparisons with existing methods. The added value of the developed method is further illustrated via the analysis of a real data example.

Keywords: B-spline approximation; Nonlinear functional modeling; Phase and amplitude variation; Random effects; Warping function.

1 Introduction

Functional data are encountered in many fields, a multitude of examples can be found in the books by Ferraty and Vieu (2006); Ramsay and Silverman (2002, 2005). When analyzing functional data it is of particular interest to provide answers to the questions: (i) is there a common main (mean) functional pattern to be distinguished?; (ii) can we quantify the significant individual fluctuations with respect to such a mean pattern? While the common functional mean is capturing main features such as peaks and valleys, differences between individual curves are often exposed via differences in phase and in amplitude of the main features. In Figure 1 the Pinch Force data are depicted. These data were collected as part of an experiment to investigate the force (measured in Newton) exerted by thumb and forefinger when pinching a 6 cm width force meter. See Ramsay et al. (1995). Data on 20 recordings of such force measurements, recorded every 2 milliseconds during a time period of 0.3 seconds, are presented in Figure 1. There seems to be a clear maximum for each curve, but the position and the size of this maximum differs considerably from curve to curve. See further Section 5.

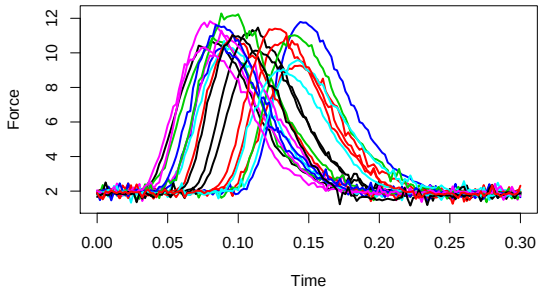

Figure 1: Pinch Force dataset.

Often there is no prior information available regarding the number of important features, and where, in which region, they occur. A flexible method should thus not rely on such information, and be able to

[1]The first and third author gratefully acknowledge support from the C1-project C16/20/002 of the KU Leuven Research Fund. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government. Part of this work was accomplished when the second author was a Postdoctoral Researcher at the KU Leuven.

extract a main pattern from the data, as well as information on major individual variations. Aligning the individual curves via individual shift functions is conveniently done via time warping functions, see for example Bigot (2013); Claeskens et al. (2010); Dupuy et al. (2011); Gervini and Gasser (2004); Kneip and Gasser (1992); Wang and Gasser (1997). One approach towards describing the curve-specific deviations from the mean curve is via random effects. See for example Chen and Wang (2011); Elmi et al. (2011); Guo (2002). An analysis of variance model for functional data describing the phase variability through time-warping and allowing for inference in the presence of amplitude variability, was introduced by Gervini and Carter (2014). This approach was further extended to a functional regression setting in Gervini (2015b,a). A functional mixed effects regression model was used to analyse spike train data in Hadjipantelis et al. (2014). In Xie et al. (2017) the emphasis was to construct separate boxplot-type displays for the three main components of the observed variation in functional data, the amplitude, phase, and vertical translation. A shift-warping method is used in Carroll et al. (2020) for multivariate functional data where each of the components may contribute to a shift with its own parameter value. In Wrobel et al. (2019), warping methods are proposed for data from exponential families. The methodology consists of working with principal components analysis (PCA) and using an expectation-maximization (EM) algorithm for parameter estimation. In Happ et al. (2019), PCA is studied to analyse warping functions. A nonparametric registration method is proposed in Chakraborty and Panaretos (2021), based on a local variation measure introduced to provide nonparametric conditions that lead to identifiability. The phase and amplitude are separated in Tucker et al. (2013) by using a representation of functional data that is based on the Fisher-Rao metric to compute an elastic shape analysis of the curves. Based on this representation, Yu et al. (2017) analyses the phase variation using a principal nested sphere approach. In Strait et al. (2017), a constrained elastic shape analysis is used with a landmark representation. While there are Bayesian methods for registration too, see for example Cheng et al. (2016), these are not considered here. Other papers focus on curve registration and classification or clustering, see Park and Ahn (2017); Sangalli et al. (2010); Tang et al. (2020); Zeng et al. (2019).

In this paper we use a mixed effects model in which random effects enter on two levels: (1) a warping function with random effects describes the individual phase variability in a flexible manner, and (2) a second random effect is used to model the individual amplitude variability. This follows the approach of Gervini and Carter (2014), but with two major differences: (i) the definition of the warping function does not depend on 'landmarks' (locations of peaks and valleys); (ii) the estimation procedure. A first important advantage of our method is that there is no need to know nor estimate landmarks, neither their number nor their positions, which can be time consuming and/or difficult. Second, our estimation procedure is computationally less demanding than, for example, an EM-algorithm as used in Gervini and Carter (2014). Different from Rakêt et al. (2014) is that we use nonparametric estimation by means of B-splines and avoid a linearization of the mean around the warped values as in their estimation approach. We focus in this paper on homogeneous signals. We prove the identifiability results of the proposed data registration model under some mild conditions. In addition, the asymptotic properties of the proposed estimation procedure are investigated: convergence of the algorithm, asymptotic normality of the estimator at each step, and consistency of the final estimator. An important contribution of this theory is the study of the algorithm, seen as an iterative process, and not on the estimator that it approximates. The added value of the method is illustrated on the Pinch Force data in Section 5, where our analysis not only provides a mean pattern, but also allows to describe clearly where most individual differences occur with respect to either phase or amplitude.

The paper is organized as follows. In Section 2 the modeling framework is introduced together with the necessary notations. The identifiability of the proposed model is obtained. Details about the estimation procedure are provided in Section 3. An estimator for the warping parameters is constructed, and its asymptotic normality is proven. Linear mixed model estimators are proposed for the functional parameters and their asymptotic normality is shown. In Section 3.3 we derive an iterative estimation procedure for which we show the convergence and the consistency of the resulting estimators. The finite-sample performance of the proposed estimation method is investigated in Section 4, which also includes a comparison with four existing methods. The methodology is used to analyse the Pinch Force data in Section 5. The paper concludes by some discussion in Section 6. This paper is accompanied by the R package warpMix. All proofs are given in the Appendices.

2 The model and its identifiability

Suppose one observes individual curves $Y_1(t), Y_2(t), \ldots, Y_n(t)$ on the interval $[0, 1]$ (without loss of generality), and a first aim is to find a main pattern $\mu(t)$ in these individual curves. First, we introduce the

various elements of the modeling framework, and provide the identifiability of the model. All the proofs of the results stated in this section can be found in Appendix A.

2.1 A functional mixed model with warping function

We consider the following functional mixed model. For $i = 1, \dots, n$, and for $t \in [0,1]$, we define the process

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

with μ the unknown common mean and where U_i denotes the unknown random effect on the amplitude for the observation i. The flexible warping function $w : [0,1] \to [0,1]$ is strictly increasing and depends on a random variable $\boldsymbol{\theta}_i \sim \mathcal{N}_r(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$, that describes the individual phase variability. Details about the warping function are provided in Section 2.3.

We rather use the discretization of model (1) with time points $(t_{i,j})$ for $j = 1, \dots, T_i$; $i = 1, \dots, n$, where T_i denotes the number of fixed (non-random) time points for the individual i:

$$Y_i(t_{i,j}) = \mu\left\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i)\right\} + U_i\left\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i)\right\} + \varepsilon_{i,j}. \tag{2}$$

We assume that for all i, the error vectors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \dots, \varepsilon_{i,T_i})^\top$ with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}(\mathbf{0}_{T_i}, \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_{T_i})$ are i.i.d., meaning that the error terms are independent of $t_{i,j}$ and of the warping effects $\boldsymbol{\theta}_i$.

An analogous model was used in Rakêt et al. (2014), where the warping function stands only in the common mean and not in the individual effect, and in Gervini and Carter (2014), where a group level is added. We argue that from this general formulation, many things have to be defined to allow the estimation of this model. The specificities of our study and its novelty will be described in the next two subsections, through the decomposition of the signals onto a B-splines basis and the warping function.

2.2 B-spline basis decomposition

The warping function w, the unknown mean function μ and the individual random effect amplitude functions U_i are modeled in a flexible fashion via B-splines. In this paper, we make the following assumption.

Assumption A. *We assume that the functions μ, $(U_i)_{i=1,\dots,n}$ and w belong to the space spanned by the considered spline basis.*

Assumption A ensures to have unbiased estimators for the curves, and avoids having to theoretically deal with a modeling bias. When using a spline basis in practice, the curves are well approximated when utilizing a finite (maybe large) number of knots.

For the mean function μ, we define a sequence of K_μ interior knots $0 = \kappa^\mu_0 < \kappa^\mu_1 < \dots \kappa^\mu_{K_\mu} < \kappa^\mu_{K_\mu+1} = 1$. In addition, we put $p_\mu + 1$ boundary knots at 0 as well as at 1, and denote $\kappa^\mu_{-p_\mu} = \dots = \kappa^\mu_{-1} = \kappa^\mu_0$ and $\kappa^\mu_{K_\mu+1} = \kappa^\mu_{K_\mu+2} = \dots = \kappa^\mu_{K_\mu+p_\mu+1}$. We denote by $\boldsymbol{\kappa}^\mu = \{\kappa^\mu_{-p_\mu}, \dots, \kappa^\mu_{K_\mu+p_\mu+1}\}$ the set of all knots involved in estimation of μ. The B-spline basis functions of degree p_μ are defined by induction as

$$B^\mu_{l,1}(t;\boldsymbol{\kappa}^\mu) = \begin{cases} 1 & \text{if } \kappa^\mu_l \le t \le \kappa^\mu_{l+1}; \\ 0 & \text{otherwise;} \end{cases}$$

$$B^\mu_{l,p_\mu+1}(t;\boldsymbol{\kappa}^\mu) = \frac{t-\kappa^\mu_l}{\kappa^\mu_{l+p_\mu}-\kappa^\mu_l} B^\mu_{l,p_\mu}(t;\boldsymbol{\kappa}^\mu) + \frac{\kappa^\mu_{l+p_\mu+1}-t}{\kappa^\mu_{l+p_\mu+1}-\kappa^\mu_{l+1}} B^\mu_{l+1,p_\mu}(t;\boldsymbol{\kappa}^\mu);$$

for $l = -p_\mu, \dots, K_\mu$. With the use of the additional (boundary) knots, this gives precisely $m_\mu = K_\mu + p_\mu + 1$ basis functions. The function μ is decomposed in the B-spline basis, with coefficient vector $\boldsymbol{\alpha}^\mu = (\alpha^\mu_{-p_\mu}, \dots, \alpha^\mu_{K_\mu})^\top$,

$$\mu(t) = \sum_{l=-p_\mu}^{K_\mu} \alpha^\mu_l B^\mu_{l,p_\mu+1}(t;\boldsymbol{\kappa}^\mu). \tag{3}$$

Note that if $\mu(\cdot)$ does not belong to the space spanned by the basis functions, then the equality in (3) should be replaced by an approximation. The induced modeling bias can be controlled by taking a large number of knots.

Similarly, for $i = 1, \dots, n$ each individual random function U_i is decomposed in a basis of B-splines of degree p_{U_i}. Denote the B-spline basis for U_i by $(B^U_{i,-p_{U_i},p_{U_i}+1}, \dots, B^U_{i,K_{U_i},p_{U_i}+1})$ with knots sequence $\boldsymbol{\kappa}^{U_i}$, resulting in $m_{U_i} = K_{U_i} + p_{U_i} + 1$ basis functions, and consider

$$U_i(t) = \sum_{l=-p_{U_i}}^{K_{U_i}} \alpha^U_{i,l} B^U_{i,l,p_{U_i}+1}(t; \boldsymbol{\kappa}^{U_i}),$$

where $\boldsymbol{\alpha}^U_i = (\alpha^U_{i,-p_{U_i}}, \dots, \alpha^U_{i,K_{U_i}})^\top$. For this random vector $\boldsymbol{\alpha}^U_i$ of B-spline coefficients, attaining values in $\mathbb{R}^{m_{U_i}\times 1}$, we assume that, for all $i = 1, \dots, n$,

$$\boldsymbol{\alpha}^U_i = \begin{pmatrix} \alpha^U_{i,-p_{U_i}} \\ \vdots \\ \alpha^U_{i,K_{U_i}} \end{pmatrix} \sim \mathcal{N}_{m_{U_i}}\left(\mathbf{0}_{m_{U_i}}, \Sigma^{U_i}\right) \text{ with } \Sigma^{U_i} = \begin{pmatrix} \sigma^2_{U,1} & 0 & \dots & 0 \\ 0 & \sigma^2_{U,2} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2_{U,m_{U_i}} \end{pmatrix},$$

the covariance matrix for which we assume a diagonal structure, and which needs to be estimated. Further we denote $\underset{\sim}{\boldsymbol{\alpha}}^U = \left((\boldsymbol{\alpha}^U_1)^\top, \dots, (\boldsymbol{\alpha}^U_n)^\top\right)^\top$, a random vector taking values in $\mathbb{R}^{\sum_{i=1}^n m_{U_i}\times 1}$.

2.3 The warping function

A flexible way to model the warping function is as follows. For every $t \in [0,1]$, we define

$$w^{-1}(t; \boldsymbol{\theta}_i) = \frac{\int_0^t \exp\left\{h^{-1}(u; \boldsymbol{\theta}_i)\right\} du}{\int_0^1 \exp\left\{h^{-1}(u; \boldsymbol{\theta}_i)\right\} du}, \tag{4}$$

with h^{-1} as indicated below. Note that w^{-1} (and hence w) is by construction an increasing function. The advantage of using the exponential function is that it warrants the positivity of the function. A non-random version of this warping function was introduced in Ramsay and Silverman (2005) and used in Hadjipantelis et al. (2014); Wagner and Kneip (2019). There are many other choices of warping functions that could be made (see for example Marron et al. (2015)). In short, the warping function w^{-1} (or w) in (4) satisfies the following necessary conditions: increasing, and from $[0,1]$ to $[0,1]$. To ensure identifiability, the function h^{-1} will be decomposed using a basis of centralized B-splines, i.e.

$$h^{-1}(u; \boldsymbol{\theta}_i) = \sum_{l=-p_h}^{K_h} \theta_{i,l} \bar{B}^h_{l,p_h+1}(u; \boldsymbol{\kappa}^h), \tag{5}$$

where $(\bar{B}^h_{l,p_h+1})_{l=-p_h,\dots,K_h}$ satisfy

$$\int_0^1 \bar{B}^h_{l,p_h+1}(u; \boldsymbol{\kappa}^h) du = 0.$$

The vector of random effects $\boldsymbol{\theta}_i = (\theta_{i,-p_h}, \dots, \theta_{i,K_h})^\top$ describes the individual phase variability, for which we assume a linear mixed effects model

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \mathbf{E}_i + \tilde{\boldsymbol{\varepsilon}}_i, \tag{6}$$

with $\mathbf{E}_i \sim \mathcal{N}_r(\mathbf{0}_r, \Sigma^{\mathbf{E}})$ and $\tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}_r(\mathbf{0}_r, \sigma^2_{\tilde{\boldsymbol{\varepsilon}}}\mathbf{I}_r)$ independent. Then $\boldsymbol{\theta}_i \sim \mathcal{N}_r(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$, with $\Sigma^{\boldsymbol{\theta}} = \Sigma^{\mathbf{E}} + \sigma^2_{\tilde{\boldsymbol{\varepsilon}}}\mathbf{I}_r$. To ensure identifiability, we assume that $\sigma^2_{\tilde{\boldsymbol{\varepsilon}}}$ is known. In Gervini and Carter (2014) the parameter $\boldsymbol{\theta}_0$ is considered to be a Jupp transform of the landmarks of the mean function μ. In contrast, we avoid the use of landmarks, and $\boldsymbol{\theta}_0$ is a parameter to be estimated.

By construction, the warping function is injective, as proved in Lemma 1.

Lemma 1. *The warping function w defined via (4) and (5) is injective with respect to the second parameter: for every $t \in [0,1]$,*

$$t = w(w^{-1}(t; \boldsymbol{\theta}^1); \boldsymbol{\theta}^2) \qquad \Rightarrow \qquad \boldsymbol{\theta}^1 = \boldsymbol{\theta}^2.$$

Further, we assume that the $\boldsymbol{\alpha}^U_i$s and $\boldsymbol{\theta}_i$s, the random effects describing the individual phase and amplitude variability, are independent of each other.

2.4 The model in matrix form

For further analysis it will be useful to introduce some matrix notation. The matrix $\mathbf{B}_i^\mu$ of dimension $T_i \times m_\mu$ contains (j,l)th element $B^\mu_{l,p_\mu+1}(t_{ij};\boldsymbol{\kappa}^\mu)$, and $\mathbf{B}_i^U$ is the matrix of dimension $T_i \times m_{U_i}$ with (j,l)th element $B^U_{i,l,p_{U_i}+1}(t_{ij};\boldsymbol{\kappa}^{U_i})$. Further, $(\mathbf{B}_i^\mu)^{\boldsymbol{\theta}_i} = ([(\mathbf{B}_i^\mu)^{\boldsymbol{\theta}_i}]_{j,l})_{j=1,\dots,T_i;\, l=-p_\mu,\dots,K_\mu}$, with

$$[(\mathbf{B}_i^\mu)^{\boldsymbol{\theta}_i}]_{j,l} = B^\mu_{l,p_\mu+1}\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i);\boldsymbol{\kappa}^\mu\}.$$

Define $[(\mathbf{B}_i^\mu)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}]_{j,l} = B^\mu_{l,p_\mu+1}[w^{-1}\{w(t_{i,j};\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\};\boldsymbol{\kappa}^\mu]$ for $j = 1,\dots,T_i$ and $l = -p_\mu,\dots,K_\mu$. Similarly, we define $(\mathbf{B}_i^U)^{\boldsymbol{\theta}_i}$ and $[(\mathbf{B}_i^U)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}]_{j,l}$.

The model (2) in matrix representation is

$$\underset{\sim}{\mathbf{Y}} = \underset{\sim}{\mathbf{B}}^\mu \boldsymbol{\alpha}^\mu + \underset{\sim}{\mathbf{B}}^U \underset{\sim}{\boldsymbol{\alpha}}^U + \underset{\sim}{\boldsymbol{\varepsilon}} \tag{7}$$

with

$$\begin{aligned}
\underset{\sim}{\mathbf{Y}} &= \left((Y_1(t_{1,1}),\dots,Y_1(t_{1,T_1}))^\top,\dots,(Y_n(t_{n,1}),\dots,Y_n(t_{n,T_n}))^\top\right)^\top \in \mathbb{R}^{\sum_{i=1}^n T_i \times 1},\\
\underset{\sim}{\boldsymbol{\varepsilon}} &= \left((\boldsymbol{\varepsilon}_1)^\top,\dots,(\boldsymbol{\varepsilon}_n)^\top\right)^\top \sim \mathcal{N}_{\sum_{i=1}^n T_i}(\mathbf{0}_{\sum_{i=1}^n T_i}, \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_{\sum_{i=1}^n T_i}),\\
\boldsymbol{\alpha}^\mu &\in \mathbb{R}^{m_\mu \times 1},\\
\underset{\sim}{\mathbf{B}}^\mu &= \left[(\mathbf{B}_1^\mu)^{\boldsymbol{\theta}_1};\dots;(\mathbf{B}_n^\mu)^{\boldsymbol{\theta}_n}\right] \in \mathbb{R}^{\sum_{i=1}^n T_i \times m_\mu},\\
\underset{\sim}{\boldsymbol{\alpha}}^U &\sim \mathcal{N}_{\sum_{i=1}^n m_{U_i}}(\mathbf{0}_{\sum_{i=1}^n m_{U_i}}, \underset{\sim}{\Sigma}^U),\\
\underset{\sim}{\Sigma}^U &= \begin{pmatrix} \Sigma^{U_1} & 0 & 0 & 0\\ 0 & \Sigma^{U_2} & 0 & 0\\ & & & \\ 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & \Sigma^{U_n}\end{pmatrix},\\
\underset{\sim}{\mathbf{B}}^U &= \left[(\mathbf{B}_1^U)^{\boldsymbol{\theta}_1};\dots;(\mathbf{B}_n^U)^{\boldsymbol{\theta}_n}\right] \in \mathbb{R}^{\sum_{i=1}^n T_i \times m_\mu},\\
\underset{\sim}{\boldsymbol{\theta}} &= \left((\boldsymbol{\theta}_1)^\top,\dots,(\boldsymbol{\theta}_n)^\top\right)^\top \sim \mathcal{N}_{r\times n}(\underset{\sim}{\boldsymbol{\theta}}_0, \mathbf{I}_r \otimes \Sigma^{\boldsymbol{\theta}}),\\
\underset{\sim}{\boldsymbol{\theta}}_0 &\in \mathbb{R}^{(r\times n)\times 1},
\end{aligned}$$

where $\mathbf{A}\otimes\mathbf{B}$ denotes the Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$.

In summary, the unknown parts in the model consist of $\boldsymbol{\alpha}^\mu, \sigma^2_{\boldsymbol{\varepsilon}}, \underset{\sim}{\Sigma}^U, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}$.

2.5 Identifiability of the model

In this section, we provide sufficient and necessary conditions to ensure the identifiability of model (7). First, the joint model (7) is identifiable if and only if at least one (approximate) individual model (2) is identifiable. We thus focus on a fixed i, and on the set of parameters $(\boldsymbol{\alpha}^\mu, \sigma^2_{\boldsymbol{\varepsilon}}, \Sigma^{U_i}, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ which consists of the subparts $(\boldsymbol{\alpha}^\mu, \sigma^2_{\boldsymbol{\varepsilon}}, \Sigma^{U_i})$ and $(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$, where the latter is linked to the warping modeling part, and the former with the other parts. We start by investigating identifiability in each part.

2.5.1 Identifiability of the warped process

Take any $i \in \{1,\dots,n\}$. For $j = 1,\dots,T_i$, let $\mathbf{X}_i$ be the warped process:

$$\begin{aligned}
X_{i,j} &= Y_i\{w(t_{i,j};\boldsymbol{\theta}_i)\} = \mu(t_{i,j}) + U_i(t_{i,j}) + \varepsilon_{i,j}\\
&= \sum_{l=-p_\mu}^{K_\mu} \alpha^\mu_l B^\mu_{l,p_\mu+1}(t_{i,j};\boldsymbol{\kappa}^\mu) + \sum_{l=-p_{U_i}}^{K_{U_i}} \alpha^U_{i,l} B^U_{i,l,p_{U_i}+1}(t_{i,j};\boldsymbol{\kappa}^{U_i}) + \varepsilon_{i,j}.
\end{aligned} \tag{8}$$

Since $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}(\mathbf{0}_{T_i}, \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_{T_i})$, we obtain

$$\boldsymbol{X}_i|\boldsymbol{\alpha}_i^U \sim \mathcal{N}_{T_i}(\mathbf{B}_i^\mu\boldsymbol{\alpha}^\mu + \mathbf{B}_i^U\boldsymbol{\alpha}_i^U, \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_{T_i}).$$

First, remark that if we know $(\sigma^2_{\boldsymbol{\varepsilon}}, \Sigma^{U_i})$, or if we know $\sigma^2_{\boldsymbol{\varepsilon}}$, or if we know Σ^{U_i}, model (8) is identifiable. In the following theorem, we give sufficient and necessary conditions for model (8) to be identifiable when both variance parameters are unknown. Since, for given $(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$, and due to the use of B-spline approximations, the warped process leads to a linear mixed effects model, we can use general results on the identifiability of such models, as obtained by Wang (2013). Theorem 1 is an adaptation of Corollary 4.2 in Wang (2013) to the current setting.

Theorem 1. *Let $i \in \{1, \dots, n\}$ be given. Model (8) is not identifiable if and only if the three conditions are fulfilled:*

1. $(\mathbf{B}_i^U)^\top \mathbf{B}_i^U \neq \mathbf{0}_{m_{U_i}}$;
2. $\mathbf{H}_i^U = \mathbf{B}_i^U \{(\mathbf{B}_i^U)^\top \mathbf{B}_i^U\}^{-1} (\mathbf{B}_i^U)^\top = \mathbf{I}_{T_i}$;
3. $(\mathbf{B}_i^U)^\top \mathbf{B}_i^U$ *is diagonal.*

Consequently, model (8) is identifiable if at least one of the three conditions in Theorem 1 is not satisfied.

2.5.2 Identifiability of the warping function

Here, we assume that we know the parameters of the mixed effects model $(\boldsymbol{\alpha}^\mu, \sigma^2_{\boldsymbol{\varepsilon}}, \Sigma^{U_i})$, and we want to prove the identifiability of the warping process

$$Y_i(t) = X_i\{w^{-1}(t; \boldsymbol{\theta}_i)\}, \tag{9}$$

involving the parameters $(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$.

Sufficient and necessary conditions for identifiability of this part are established in Theorem 2.

Theorem 2. *Let $i \in \{1, \dots, n\}$ be given. Let $\boldsymbol{\theta}_i \sim \mathcal{N}_r(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}}_i \sim \mathcal{N}_r(\tilde{\boldsymbol{\theta}}_0, \Sigma^{\tilde{\boldsymbol{\theta}}})$ be used to define two warping functions $w^{-1}(.; \boldsymbol{\theta}_i)$ and $w^{-1}(.; \tilde{\boldsymbol{\theta}}_i)$, and let X_i and $\tilde{X}_i$ be the corresponding warped functions, such that*

$$Y_i(t) = X_i\{w^{-1}(t; \boldsymbol{\theta}_i)\} = \tilde{X}_i\{w^{-1}(t; \tilde{\boldsymbol{\theta}}_i)\}.$$

Then model (9) is identifiable if and only if

$$\mathbf{B}_i^\mu = \mathrm{E}_{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i} \left\{ (\mathbf{B}_i^\mu)^{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i} \right\};$$

$$(\mathbf{B}_i^U)^\top \Sigma^{U_i} \mathbf{B}_i^U = \mathrm{Var}_{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i} \left\{ (\mathbf{B}_i^\mu)^{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i} \boldsymbol{\alpha}^\mu \right\} + \mathrm{E}_{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i} \left[\left\{ (\mathbf{B}_i^U)^{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i} \right\}^\top \Sigma^{U_i} (\mathbf{B}_i^U)^{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i} \right].$$

2.5.3 Identifiability of the global model

We proved that, when knowing the warping parameters, the functional linear mixed effects model is identifiable, and that when knowing the functional linear mixed effects model, the warping parameters are identifiable. Then, by iterating between these two identifiable steps until convergence, we have a procedure which is identifiable and leads to the estimation of all the parameters of the model defined in (2). Note that the identifiability conditions are essentially conditions on the englobing B-spline basis structure.

3 Estimation procedure and asymptotic properties

Recall that the unknown parameters of model (2) are $(\boldsymbol{\alpha}^\mu, \sigma^2_{\boldsymbol{\varepsilon}}, \underset{\sim}{\Sigma}^U, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$. Model (2) is a *nonlinear* functional mixed effects model due to the composition by the warping function, which is an essential ingredient to describe the individual phase variability. First, we analyse each part of the model, that is, the warping parameters and the linear mixed effect model, by providing an estimator and theoretical guarantees. Then, we propose an iterative estimation procedure, where in a first step we fix the warping parameters $(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ and estimate the functional parameters $(\boldsymbol{\alpha}^\mu, \sigma^2_{\boldsymbol{\varepsilon}}, \underset{\sim}{\Sigma}^U)$; and next, we start from these estimated parameters, and estimate the warping parameters. Further, we obtain the convergence of the method and the consistency of the global estimator.

We have access to a sample $(Y_i(t_{i,j}))_{j=1,\dots,T_i;\, i=1,\dots,n}$ of n curves, the ith curve being evaluated in T_i points. Given are the knot sequences in the B-splines approximations, the degree of the B-splines, and the dimension of the warping parameters $r = K_h + p_h + 1$.

Proofs of the results in this section are provided in Appendix B.

3.1 Parameters of the warping function

3.1.1 Estimators for the parameters of the warping function

Suppose we know the functional parameters $(\boldsymbol{\alpha}^{\mu}, \underset{\sim}{\Sigma}^{U}, \sigma^2_{\boldsymbol{\varepsilon}})$, and the predictors $\boldsymbol{\alpha}_i^U$ for all $i = 1, \dots, n$. The goal is to estimate $(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$.

We construct pseudo-observations by minimizing the following empirical L_2 criterion:

$$\hat{\boldsymbol{\theta}}_i^{T_i} = \underset{\tilde{\boldsymbol{\theta}}_i \in \mathbb{R}^r}{\operatorname{argmin}} \left[\sum_{j=1}^{T_i - 1} \left\{ Y_i\{w(t_{i,j}; \tilde{\boldsymbol{\theta}}_i)\} - \mu(t_{i,j}) - U_i(t_{i,j}) \right\}^2 (t_{i,j+1} - t_{i,j}) \right]. \tag{10}$$

Note that the criterion which is minimized tends to the L_2 distance between $Y_i \circ w(.; \tilde{\boldsymbol{\theta}}_i)$ and $\mu + U_i$, if $T_i \to +\infty$.

However, as we want to consider the warping parameter as a random effect, we fit a mixed effects model as defined in Eq. (6) on the pseudo-observations $\hat{\boldsymbol{\theta}}_1^{T_1}, \dots, \hat{\boldsymbol{\theta}}_n^{T_n}$, that is,

$$\hat{\boldsymbol{\theta}}_i^{T_i} = \boldsymbol{\theta}_0 + \mathbf{E}_i + \tilde{\boldsymbol{\varepsilon}}_i, \tag{11}$$

with $\mathbf{E}_i \sim \mathcal{N}_r(\mathbf{0}_r, \Sigma^{\mathbf{E}})$, $\tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}_r(\mathbf{0}_r, \sigma^2_{\tilde{\boldsymbol{\varepsilon}}} \mathbf{I}_r)$ for all $i = 1, \dots, n$. The random variables $\mathbf{E}_i$ and $\tilde{\boldsymbol{\varepsilon}}_i$ are independent. As we assume that $\sigma^2_{\tilde{\boldsymbol{\varepsilon}}}$ is known for identifiability reasons, we use the empirical mean of $\{\hat{\boldsymbol{\theta}}_1^{T_1}, \dots, \hat{\boldsymbol{\theta}}_n^{T_n}\}$ to estimate $\boldsymbol{\theta}_0$, and the empirical covariance to estimate $\Sigma^{\boldsymbol{\theta}} = \Sigma^{\mathbf{E}} + \sigma^2_{\tilde{\boldsymbol{\varepsilon}}} \mathbf{I}_r$. The prediction of $\mathbf{E}_i$ is easy to get because $\sigma^2_{\tilde{\boldsymbol{\varepsilon}}}$ is known. We consider the following estimators:

$$\hat{\boldsymbol{\theta}}_0 = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_i^{T_i}; \qquad \hat{\Sigma}^{\boldsymbol{\theta}} = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_i^{T_i} - \hat{\boldsymbol{\theta}}_0)(\hat{\boldsymbol{\theta}}_i^{T_i} - \hat{\boldsymbol{\theta}}_0)^{\top};$$

$$\hat{\mathbf{E}}_i = \left(\hat{\Sigma}^{\boldsymbol{\theta}} - \sigma^2_{\tilde{\boldsymbol{\varepsilon}}} \mathbf{I}_r \right) \left(\hat{\Sigma}^{\boldsymbol{\theta}} \right)^{-1} \hat{\boldsymbol{\theta}}_i^{T_i}.$$

3.1.2 Asymptotic normality of $\hat{\boldsymbol{\theta}}_i^{T_i}$

First, we focus on the distribution of $\hat{\boldsymbol{\theta}}_i^{T_i}$ conditional on $\boldsymbol{\theta}_i$. To do so, we rely on the theory of nonlinear least squares estimators developed in Jennrich (1969), which uses the weighted tail product defined as follows.

Definition 1. *Let p be a nonnegative integer and $(t_j)_{j=1,\dots,p}$ be fixed time points. Let $x = (x_p)$ and $y = (y_p)$ be two sequences of real numbers and let*

$$(x, y)_p^{\pi} = \frac{1}{p} \sum_{j=1}^{p-1} x_j y_j (t_{j+1} - t_j).$$

If $(x, y)_p^{\pi}$ converges to a real number when $p \to +\infty$, its limit $(x, y)^{\pi}$ is called the weighted tail product *of x and y.*

Let g and h be two sequence valued functions on Θ. If $(g(\alpha), h(\beta))_p^{\pi} \to (g(\alpha), h(\beta))^{\pi}$ when $p \to +\infty$ uniformly for all α and β in Θ, we define

$$[g, h] : (\alpha, \beta) \in \Theta \times \Theta \mapsto (g(\alpha), h(\beta))^{\pi}.$$

This function is called the weighted tail cross product *of g and h.*

Then, we define the $r \times r$ -matrix $\mathbf{a}_i(\tilde{\boldsymbol{\theta}}_i)$ as follows.

Definition 2. *For $l = 1, \dots, r$, we denote by*

$$\partial_l(\mu + U_i) = \frac{\partial[(\mu(w^{-1}(t_{i,j}; \tilde{\boldsymbol{\theta}}_i)) + U_i(w^{-1}(t_{i,j}; \tilde{\boldsymbol{\theta}}_i)))_{j=1,\dots,T_i}]}{\partial[\tilde{\boldsymbol{\theta}}_i]_l}.$$

the partial derivative of the aligned signal. We define

$$\mathbf{a}_{i,T_i}(\tilde{\boldsymbol{\theta}}_i) = \left[(\partial_l(\mu + U_i), \partial_{l'}(\mu + U_i))_{T_i}^{\pi} \right]_{l=1,\dots,r; l'=1,\dots,r}$$

the matrix with coefficients the weighted tail product between two partial derivatives, and $\mathbf{a}_i(\tilde{\boldsymbol{\theta}}_i)$ its limit when $T_i \to +\infty$.

Assumption B. *For all $i = 1, \dots, n$, the $r \times r$-matrix $\mathbf{a}_i(\boldsymbol{\theta}_i)$ is non-singular.*

Theorem 3. *Fix $i \in \{1, \dots, n\}$, and $\alpha \in (0,1)$. Let $(\hat{\boldsymbol{\theta}}_i^{T_i})_{T_i}$ be a sequence of weighted least squares estimators of $\boldsymbol{\theta}_i$. We assume that the model is identifiable and satisfies Assumption A. With probability $1-\alpha$, conditional on $\boldsymbol{\theta}_i$, $\hat{\boldsymbol{\theta}}_i^{T_i}$ is a strongly consistent estimator of $\boldsymbol{\theta}_i$ (convergence a.s.).*

If we assume that the model satisfies Assumption B, conditional on $\boldsymbol{\theta}_i$,

$$T_i^{1/2}(\hat{\boldsymbol{\theta}}_i^{T_i} - \boldsymbol{\theta}_i) \xrightarrow[T_i \to +\infty]{d} \mathcal{N}_r(0, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_i));$$

and $\mathbf{a}_{i,T_i}(\hat{\boldsymbol{\theta}}_i^{T_i})$ is a strongly consistent estimator of $\mathbf{a}_i(\boldsymbol{\theta}_i)$.

Let $f_{\hat{\boldsymbol{\theta}}_i^{T_i}}(.|\boldsymbol{\theta}_i)$ be the conditional distribution function. Denote by φ the Gaussian density function. Theorem 3 implies that for all $\eta \in \mathbb{R}$,

$$f_{\hat{\boldsymbol{\theta}}_i^{T_i}}(\eta|\boldsymbol{\theta}_i) \underset{T_i \to +\infty}{\to} \varphi(\eta; \boldsymbol{\theta}_i, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_i)).$$

By the dominated convergence theorem, we get the asymptotic marginal distribution, for all η:

$$\begin{aligned} m_{\hat{\boldsymbol{\theta}}_i^{T_i}}(\eta) \underset{T_i \to +\infty}{\to} m_{\hat{\boldsymbol{\theta}}_i^{\infty}}(\eta) &= \int \varphi(\eta; \boldsymbol{\theta}_i, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_i)) \varphi(\boldsymbol{\theta}_i; \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}) d\boldsymbol{\theta}_i \\ &= \int_{\mathbb{R}^r} \varphi(\boldsymbol{\theta}_i; c, C) \varphi(\boldsymbol{\theta}_0; \eta, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) + \Sigma^{\boldsymbol{\theta}}) d\boldsymbol{\theta}_i \end{aligned}$$

where $\hat{\boldsymbol{\theta}}_i^{\infty} = \lim_{T_i \to \infty} \hat{\boldsymbol{\theta}}_i^{T_i}$, and

$$\begin{aligned} C &= (\sigma_\varepsilon^{-2} \mathbf{a}_i(\boldsymbol{\theta}_i) + (\Sigma^{\boldsymbol{\theta}})^{-1})^{-1}; \\ c &= \left(\sigma_\varepsilon^{-2} \mathbf{a}_i(\boldsymbol{\theta}_i) + (\Sigma^{\boldsymbol{\theta}})^{-1}\right)^{-1} \left(\sigma_\varepsilon^{-2} \mathbf{a}_i(\boldsymbol{\theta}_i) \hat{\boldsymbol{\theta}}_i^{\infty} + (\Sigma^{\boldsymbol{\theta}})^{-1} \boldsymbol{\theta}_0\right). \end{aligned}$$

The last line comes from a computation with Gaussian densities, see Lemma 4 in Appendix C.

We discuss two cases where the limiting distribution $m_{\hat{\boldsymbol{\theta}}_i^{\infty}}$ is computed explicitly:

- the case when the noise ε tends to disappear, which makes the theory easier, but also requires a strong assumption for the limit to hold;
- the case when the eigenvalues of the matrix $\mathbf{a}_i(\boldsymbol{\theta}_i)$ are bounded. Under this weak assumption the limiting distribution will be more complicated (see below).

We next discuss these two cases in more detail.

In the first case we assume that the noise tends to disappear, when the number of points in the time grid increases.

Assumption C. *We assume that $\sigma_\varepsilon \underset{\min T_i \to \infty}{\to} 0$.*

It is important to remark the following. If, however, there is a non-negligible noise, the method will warp the observed noise curve on some global mean, and the warping parameter will depend on this noise, whereas the true warping parameter does not, as it would be based on the denoised data.

Theorem 4. *Fix $i \in \{1, \dots, n\}$. Let $(\hat{\boldsymbol{\theta}}_i^{T_i})_{T_i}$ be a sequence of weighted least squares estimators of $\boldsymbol{\theta}_i$, and $\hat{\boldsymbol{\theta}}_i^{\infty} = \lim_{T_i \to \infty} \hat{\boldsymbol{\theta}}_i^{T_i}$. We assume that the model is identifiable and satisfies Assumptions A, B and C. Then, for all $\eta \in \mathbb{R}$,*

$$\lim_{\sigma_\varepsilon^2 \to 0} m_{\hat{\boldsymbol{\theta}}_i^{\infty}}(\eta) = \varphi(\eta; \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}).$$

We now turn to the second case. With weaker assumptions, the limiting distribution $m_{\hat{\boldsymbol{\theta}}_i^{\infty}}(\eta)$ is more complicate to describe. Denote by $\mathcal{E}_\alpha(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ the following ellipsoid:

$$\mathcal{E}_\rho(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}) = \left\{x \in \mathbb{R}^r \mid (\boldsymbol{\theta}_0 - x)^t (\Sigma^{\boldsymbol{\theta}})^{-1} (\boldsymbol{\theta}_0 - x) \leq \rho\right\}.$$

Assumption D. *For all $\boldsymbol{\theta}_i$, $\mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) - \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0)$ is positive definite.*

Assumption E. *The eigenvalues of* $\mathbf{a}_i$ *are bounded: there exist* λ_m, λ_M *such that, for all* $\boldsymbol{\theta}_i$*,*

$$\lambda_m \leq \min eigen(\mathbf{a}_i(\boldsymbol{\theta}_i)) \leq \max eigen(\mathbf{a}_i(\boldsymbol{\theta}_i)) \leq \lambda_M.$$

The following theorem establishes that in this second setting, the limiting distribution $m_{\hat{\boldsymbol{\theta}}_i^\infty}(\eta)$ is close to a Gaussian distribution.

Theorem 5. *Fix* $i \in \{1, \ldots, n\}$*. Let* $(\hat{\boldsymbol{\theta}}_i^{T_i})_{T_i}$ *be a sequence of weighted least squares estimators of* $\boldsymbol{\theta}_i$*, and* $\hat{\boldsymbol{\theta}}_i^\infty = \lim_{T_i \to \infty} \hat{\boldsymbol{\theta}}_i^{T_i}$*. We assume that the model is identifiable and satisfies Assumptions A, B, D and E. Let* $\rho > 0$*, and* $\mathbf{A}$ *a positive definite matrix. Then,*

$$m_{\hat{\boldsymbol{\theta}}_i^\infty}(\eta) = (1 + O(\rho))\varphi(\eta; \boldsymbol{\theta}_0, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0) + \Sigma^{\boldsymbol{\theta}}) \text{ if } \eta \in \mathcal{E}_\rho(\boldsymbol{\theta}_0, \mathbf{A}).$$

In summary, we get that under the two settings, the distribution of $\hat{\boldsymbol{\theta}}_i^\infty$ is close to a Gaussian distribution.

3.1.3 Asymptotic normality for $\hat{\boldsymbol{\theta}}_0$ and $\hat{\Sigma}^{\boldsymbol{\theta}}$

We consider the linear mixed effect model given in Eq. (11). We assume that $\sigma_{\tilde{\boldsymbol{\varepsilon}}}^2$ is known for identifiability reasons. Remark that $\Sigma^{\boldsymbol{\theta}} = \Sigma^E + \sigma_{\tilde{\varepsilon}}^2 I_{T_i}$. We are now interested in the estimators $\hat{\boldsymbol{\theta}}_0$ and $\hat{\Sigma}^{\boldsymbol{\theta}}$.

Theorem 6. *Fix* $i \in \{1, \ldots, n\}$*. Let* $(\hat{\boldsymbol{\theta}}_i^{T_i})_{T_i}$ *be a sequence of weighted least squares estimators of* $\boldsymbol{\theta}_i$*. We assume that the model is identifiable and satisfies Assumptions A, B and C.*

Let $b_{n,\mathbf{T}} = \sum_{i=1}^n T_i^{-1}$*. Then,*

$$b_{n,\mathbf{T}}^{-1/2}\left(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0\right) \underset{n \to +\infty, \min T_i \to +\infty}{\overset{d}{\to}} \mathcal{N}_r(\mathbf{0}, \Sigma^{\boldsymbol{\theta}});$$
$$\hat{\Sigma}^{\boldsymbol{\theta}} \underset{n \to +\infty, \min T_i \to +\infty}{\overset{d}{\to}} \mathcal{W}(\Sigma^{\boldsymbol{\theta}}, n-1),$$

where $\mathcal{W}(\Sigma, p)$ *denotes the Wishart distribution with scale matrix* Σ *and* p *degrees of freedom.*

Note that this implies that T_i has to go to infinity faster than n goes to infinity, i.e. $n = o(\min T_i)$. Indeed,

$$\frac{1}{\sqrt{\sum_{i=1}^n \frac{1}{T_i}}} \geq \sqrt{\frac{\min T_i}{n}}.$$

3.2 Functional parameters

Suppose we know the warping parameters $(\boldsymbol{\theta}_i)_{i=1,\ldots,n}$. Then, we warp the observations $(Y_i(t_{i,j}))_{j=1,\ldots,T_i;\, i=1,\ldots,n}$ onto the estimated warped curves $X_{i,j} = Y_i\{w(t_{i,j}; \boldsymbol{\theta}_i\}$, and we fit a functional linear mixed model on $(\mathbf{X}_i)_{i=1,\ldots,n}$ as defined in Eq. (8) using maximum likelihood estimation, which leads to estimators $(\hat{\boldsymbol{\alpha}}^\mu, \hat{\Sigma}^U, (\hat{\sigma}_{\boldsymbol{\varepsilon}}^2))$ and predictors $(\hat{\boldsymbol{\alpha}}_i^U)_{i=1,\ldots,n}$. Following the ideas described in (Pinheiro, 1994, Chapter 3), we need the following assumption:

Assumption F. *Existence and positive definiteness of* $\mathcal{I}$*, which is the limit of minus the expected Hessian matrix of the log-likelihood function based on the model given in Eq. (8).*

Then, we get the asymptotic normality of the estimator.

Theorem 7. *Let* $(\hat{\boldsymbol{\alpha}}^\mu, \hat{\Sigma}^U, (\hat{\sigma}_{\boldsymbol{\varepsilon}}^2))$ *be a sequence of maximum likelihood estimator of the functional linear mixed model, computed over the observations* $(Y_i(t_{i,j}))_{j=1,\ldots,T_i;\, i=1,\ldots,n}$*. We assume that the model is identifiable and satisfies Assumption F. Then,*

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\alpha}}^\mu - \boldsymbol{\alpha}^\mu \\ \hat{\boldsymbol{\sigma}}_U - \boldsymbol{\sigma}_U \\ \hat{\sigma}_\varepsilon - \sigma_\varepsilon \end{pmatrix} \underset{n \to \infty}{\overset{d}{\to}} \mathcal{N}(0, \mathcal{I}^{-1}).$$

3.3 Global model and iterative estimation procedure

We propose to directly estimate the nonlinear model. Working with the L_2-distance, we want to fit the model of which the coefficients minimize

$$\left\| Y - \sum_{l=-p_\mu}^{K_\mu} \alpha_l^\mu B_{l,p_\mu+1}^\mu\{w^{-1}(.;\boldsymbol{\theta});\boldsymbol{\kappa}^\mu\} - \sum_{l=-p_U}^{K_{U_i}} \alpha_l^U B_{l,p_{U_i}+1}^U\{w^{-1}(.;\boldsymbol{\theta});\boldsymbol{\kappa}^{U_i}\} \right\|_2 .$$

Using the steps described previously, we propose an iterative process that approximates the following minimizer:

$$\underset{\boldsymbol{\alpha}^\mu,\boldsymbol{\alpha}^U,(\boldsymbol{\theta}_i)_{i=1,\dots,n}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{T_i-1} \left(Y_i(t_{i,j}) - \sum_{l=-p_\mu}^{K_\mu} \alpha_l^\mu B_{l,p_\mu+1}^\mu\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i);\boldsymbol{\kappa}^\mu\} - \sum_{l=-p_{U_i}}^{K_{U_i}} \alpha_{i,l}^U B_{i,l,p_{U_i}+1}^U\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i);\boldsymbol{\kappa}^{U_i}\} \right)^2 (t_{i,j+1}-t_{i,j}) \right\}.$$

Algorithm 1 presents the steps in the iterative procedure. Further details are provided regarded the initialization, the convergence criterion, the theoretical convergence and the consistency of the resulting estimator.

Algorithm 1 WarpMix

Initialization: Computation of the deepest function $\hat{\mu}^{(0)} = \mu^{\text{deep}}$.

for $i = 1, \dots, n$ **do**

Approximation of $\boldsymbol{\theta}_i^{(0)}$ by

$$\boldsymbol{\theta}_i^{(0)} = \underset{\boldsymbol{\theta}_i\in\mathbb{R}^r}{\operatorname{argmin}} \left[\sum_{j=1}^{T_i-1} \left\{ Y_i\{w(t_{i,j};\boldsymbol{\theta}_i)\} - \hat{\mu}^{(0)}(t_{i,j}) \right\}^2 (t_{i,j+1}-t_{i,j}) \right].$$

Fit a linear mixed model on the pseudo-observations, $\boldsymbol{\theta}_i^{(0)} = \boldsymbol{\theta}_0^{(0)} + \mathbf{E}_i^{(0)} + \tilde{\boldsymbol{\varepsilon}}_i$; and deduce $\hat{\boldsymbol{\theta}}_0^{(0)}$, $(\hat{\Sigma}^{\boldsymbol{\theta}})^{(0)}$ and $\hat{\boldsymbol{\theta}}_i^{(0)} = \hat{\boldsymbol{\theta}}_0^{(0)} + \hat{\mathbf{E}}_i^{(0)}$.

for ite $= 1, \dots$ until convergence **do**

Warp the observed curves according to $w^{-1}_{\hat{\boldsymbol{\theta}}_i^{(\text{ite}-1)}}$;

Estimate $\{(\hat{\boldsymbol{\alpha}}^\mu)^{(\text{ite})}, (\hat{\boldsymbol{\alpha}}_i^U)^{(\text{ite})}, (\hat{\Sigma}^{U_i})^{(\text{ite})}, (\hat{\sigma}^2_{\boldsymbol{\varepsilon}})^{(\text{ite})}\}$ with the R package `nlme`;

Approximate $(\boldsymbol{\theta}_1^{(\text{ite})}, \dots, \boldsymbol{\theta}_n^{(\text{ite})})$ by computing, for every $i = 1, \dots, n$,

$$\boldsymbol{\theta}_i^{(\text{ite})} = \underset{\boldsymbol{\theta}_i\in\mathbb{R}^r}{\operatorname{argmin}} \left[\sum_{j=1}^{T_i-1} \left\{ Y_i\{w(t_{i,j};\boldsymbol{\theta}_i)\} - \mu(t_{i,j}) - U_i(t_{i,j}) \right\}^2 (t_{i,j+1}-t_{i,j}) \right].$$

Fit a linear mixed model on these observations: $\boldsymbol{\theta}_i^{(\text{ite})} = \hat{\boldsymbol{\theta}}_0^{(\text{ite})} + \hat{\mathbf{E}}_i^{(\text{ite})} + \tilde{\boldsymbol{\varepsilon}}_i$; and define $\hat{\boldsymbol{\theta}}_i^{(\text{ite})} = \hat{\boldsymbol{\theta}}_0^{(\text{ite})} + \hat{\mathbf{E}}_i^{(\text{ite})}$.

3.3.1 Details about the initialization

First, we initialize the mean function μ. There exist several ways to define a central curve in functional data analysis. Here we use band depth for functional data as introduced in Sun and Genton (2011), and compare every observed curve with the deepest function $\hat{\mu}^{(0)} = \mu^{\text{deep}}$. We then deduce $\hat{\boldsymbol{\alpha}}_\mu^{(0)}$, the projection of the function $\hat{\mu}^{(0)}$ onto $\mathbf{B}^\mu$ the B-spline basis considered. In the initialization step, we do not consider individual amplitude effects, i.e. $(\hat{\boldsymbol{\alpha}}_i^U)^{(0)} = \mathbf{0}_{m_{U_i}}$ for all $i = 1, \dots, n$.

3.3.2 Convergence of the algorithm

We define

$$C_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{T_i-1}\left(Y_i(t_{i,j}) - \hat{\mu}\{w^{-1}(t_{i,j};\hat{\boldsymbol{\theta}}_i)\} - \hat{U}_i\{w^{-1}(t_{i,j};\hat{\boldsymbol{\theta}}_i)\}\right)^2 (t_{i,j+1}-t_{i,j}).$$

The iterations are stopped when $C_n < 10^{-4}$ during five successive iteration steps.

Note first of all that the various iterations in Algorithm 1 involve three operations Ψ_1, Ψ_2 and Ψ_3, and that the update function to go from one iteration to the next is composed of three parts

$$\Psi = \Psi_3 \circ \Psi_2 \circ \Psi_1 : \mathbb{R}^{m_\mu + nm_U + 1 + nr} \to \mathbb{R}^{m_\mu + nm_U + 1 + nr}. \tag{12}$$

Herein Ψ and its components are defined as follows.

$$\begin{aligned}
\Psi : &((\boldsymbol{\alpha}^\mu)^{(\text{ite})}, (\boldsymbol{\alpha}_i^U)_{i=1,\dots,n}^{(\text{ite})}, (\sigma_\varepsilon^2)^{(\text{ite})}, (\boldsymbol{\theta}_i)_{i=1,\dots,n}^{(\text{ite})}) \\
&\qquad \mapsto ((\boldsymbol{\alpha}^\mu)^{(\text{ite}+1)}, (\boldsymbol{\alpha}_i^U)_{i=1,\dots,n}^{(\text{ite}+1)}, (\sigma_\varepsilon^2)^{(\text{ite}+1)}, (\boldsymbol{\theta}_i)_{i=1,\dots,n}^{(\text{ite}+1)}) \\
\Psi_1 : &((\boldsymbol{\alpha}^\mu)^{(\text{ite})}, (\boldsymbol{\alpha}_i^U)_{i=1,\dots,n}^{(\text{ite})}, (\sigma_\varepsilon^2)^{(\text{ite})}, (\boldsymbol{\theta}_i)_{i=1,\dots,n}^{(\text{ite})}) \\
&\qquad \mapsto ((\boldsymbol{\alpha}^\mu)^{(\text{ite})}, (\boldsymbol{\alpha}_i^U)_{i=1,\dots,n}^{(\text{ite})}, (\sigma_\varepsilon^2)^{(\text{ite})}, (\boldsymbol{\theta}_i)_{i=1,\dots,n}^{(\text{ite}+1)}) \\
\Psi_2 : &((\boldsymbol{\alpha}^\mu)^{(\text{ite})}, (\boldsymbol{\alpha}_i^U)_{i=1,\dots,n}^{(\text{ite})}, (\sigma_\varepsilon^2)^{(\text{ite})}, (\boldsymbol{\theta}_i)_{i=1,\dots,n}^{(\text{ite}+1)}) \\
&\qquad \mapsto ((\boldsymbol{\alpha}^\mu)^{(\text{ite})}, (\boldsymbol{\alpha}_i^U)_{i=1,\dots,n}^{(\text{ite})}, (\sigma_\varepsilon^2)^{(\text{ite})}, (\boldsymbol{\theta}_0^{(\text{ite}+1)} + E_i^{(\text{ite}+1)})_{i=1,\dots,n})
\end{aligned}$$

$$\begin{aligned}
\Psi_3 : &((\boldsymbol{\alpha}^\mu)^{(\text{ite})}, (\boldsymbol{\alpha}_i^U)_{i=1,\dots,n}^{(\text{ite})}, (\sigma_\varepsilon^2)^{(\text{ite})}, (\boldsymbol{\theta}_0^{(\text{ite}+1)} + E_i^{(\text{ite}+1)})_{i=1,\dots,n}) \\
&\qquad \mapsto ((\boldsymbol{\alpha}^\mu)^{(\text{ite}+1)}, (\boldsymbol{\alpha}_i^U)_{i=1,\dots,n}^{(\text{ite}+1)}, (\sigma_\varepsilon^2)^{(\text{ite}+1)}, (\boldsymbol{\theta}_i)_{i=1,\dots,n}^{(\text{ite}+1)}).
\end{aligned}$$

In Ψ_1 the vector $\boldsymbol{\theta}_i$ is updated. This is used as input for Ψ_2 where observations are denoised, through the linear model defined in (6). Then, this is used as input for Ψ_3, where $\boldsymbol{\alpha}^\mu, (\boldsymbol{\alpha}_i^U)_{i=1,\dots,n}, \sigma_\varepsilon^2$ are updated.

In Theorem 8 we prove that the algorithm is converging. A condition under which this holds is that Ψ_1 is a contraction mapping, as stated in the following assumption.

Assumption G. *There exists $k_{\Psi_1} < 1$ such that, for $(x, y) \in (\mathbb{R}^{m_\mu + nm_U + 1 + nr})^2$,*

$$\|\Psi_1(x) - \Psi_1(y)\|_2 \leq k_{\Psi_1}\|x - y\|_2.$$

Under Assumption G we show the convergence of the algorithm, seen as iterations of Ψ. We denote by $((\hat{\boldsymbol{\alpha}}^\mu)^{(\infty)}, (\hat{\sigma}_\varepsilon)^{(\infty)}, (\hat{\underset{\sim}{\Sigma}}^U)^{(\infty)}, \hat{\boldsymbol{\theta}}_0^{(\infty)}, (\hat{\Sigma}^{\boldsymbol{\theta}})^{(\infty)})$ the estimator obtained at the end of the algorithm.

Theorem 8. *Fix n and $\mathbf{T}$. Suppose $(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ is a sequence of iid random variables satisfying the functional nonlinear mixed model (1) observed on fixed time points: for $i = 1, \dots, n$, for $j = 1, \dots, T_i$, $[\mathbf{Y}_i]_j = Y_i(t_{i,j})$. Moreover, suppose that the model is identifiable and satisfies Assumption G.*

Then, $((\hat{\boldsymbol{\alpha}}^\mu)^{(\infty)}, (\hat{\sigma}_\varepsilon)^{(\infty)}, (\hat{\underset{\sim}{\Sigma}}^U)^{(\infty)}, \hat{\boldsymbol{\theta}}_0^{(\infty)}, (\hat{\Sigma}^{\boldsymbol{\theta}})^{(\infty)})$ exists and is unique, and the algorithm converges to this solution with a geometric rate with respect to the Euclidean distance.

This theorem gives the pointwise convergence of the algorithm. The randomness has not been taken into account. We rather focus on the iterations of steps. Theorem 8 relies on Assumption G, which appears as rather technical. To get some insights into this assumption, we investigate it in a specific setting in Example 1.

Example 1. *We focus on μ, do not consider U_i, and restrict the family of warping functions to translations: $w^{-1}(t;\theta_i) = \theta_i + t$. The global mean is supposed to be a linear function $\mu(t) = \alpha + \beta t$. Let $t_{i,1} = 0$ and $t_{i,T_i} = 1$ Finally, we set $\theta_0 = 0$.*

Fix i. Recalling (10), in this case we are looking for

$$\theta_i = \underset{\tilde{\theta}_i}{\text{argmin}} \left\{ \frac{1}{T_i}\sum_{j=1}^{T_i-1}(\alpha - \tilde{\alpha} + (\beta - \tilde{\beta})t_{i,j} - \tilde{\beta}\tilde{\theta}_i)^2(t_{i,j+1} - t_{i,j}) \right\}.$$

This is a polynomial function of degree 2 in $\tilde{\theta}_i$ with nonnegative coefficient of the quadratic term: there exists a unique minimizer:

$$\hat{\theta}_i = \frac{\alpha - \tilde{\alpha}}{\tilde{\beta}} + \frac{\beta - \tilde{\beta}}{\tilde{\beta}} \sum_{j=1}^{T_i - 1} t_{i,j}(t_{i,j+1} - t_{i,j}).$$

We know that the Lipschitz constant is bounded by the norm of the differential. Here, the function we consider is $(\tilde{\alpha}, \tilde{\beta}) \mapsto \tilde{\theta}$, so we compute the differential, evaluated in (α, β):

$$\|D_{\tilde{\alpha},\tilde{\beta}}\Psi_1\|_2^2 = \left(-\frac{1}{\tilde{\beta}}\right)^2 + \left(\frac{-\alpha + \tilde{\alpha}}{\tilde{\beta}^2} + \frac{-\beta}{\tilde{\beta}^2} \sum_{j=1}^{T_i - 1} t_{i,j}(t_{i,j+1} - t_{i,j})\right)^2;$$

$$\|D_{\alpha,\beta}\Psi_1\|_2^2 = \frac{1}{\beta^2}\left(1 + \{\sum_{j=1}^{T_i - 1} t_{i,j}(t_{i,j+1} - t_{i,j})\}^2\right).$$

These expressions reveal that for β small, the problem is more complicated (as one could expect). Note that in this special case Assumption G in fact leads to an assumption on β.

Example 1 also shows that, in some particular settings, Assumption G might be translated into a condition on μ and w.

3.3.3 Consistency of the estimators

To conclude, we provide the statistical consistency of the full procedure. This has the following meaning. When the sample size and the number of time points are going to infinity, the parameters estimated by the iterative process are converging almost-surely to the true parameter. Finally, the consistency is deduced for the common mean, seen as a functional parameter.

Theorem 9. *Suppose $(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ is a sequence of iid random variables satisfying the functional nonlinear mixed model (1) observed on fixed time points: for $i = 1, \dots, n$, for $j = 1, \dots, T_i$, $[\mathbf{Y}_i]_j = Y_i(t_{i,j})$. Suppose that the model is identifiable, and Assumptions A, B,C, F and G hold. Then,*

$$((\hat{\boldsymbol{\alpha}}^{\mu})^{(\infty)}, (\hat{\sigma}_{\varepsilon})^{(\infty)}, (\hat{\underset{\sim}{\Sigma}}^{U})^{(\infty)}, \hat{\boldsymbol{\theta}}_0^{(\infty)}, (\hat{\Sigma}^{\boldsymbol{\theta}})^{(\infty)}) \underset{\min T_i \to \infty}{\overset{a.s.}{\underset{n\to\infty}{\longrightarrow}}} (\boldsymbol{\alpha}^{\mu}, \sigma^2_{\boldsymbol{\varepsilon}}, \underset{\sim}{\Sigma}^{U}, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}).$$

As a consequence, we get that, from a functional viewpoint, for $\mu \in span(B^{\mu})$, if we denote $\hat{\mu} = (\hat{\boldsymbol{\alpha}}^{\mu})^{(\infty)} B^{\mu}$,

$$\|\mu - \hat{\mu}\|_{L_2[0,1]} \underset{\min T_i \to \infty}{\overset{a.s.}{\underset{n\to\infty}{\longrightarrow}}} 0.$$

4 Simulation study

We investigate the finite-sample performance of the proposed estimation method, and we compare it with four state-of-the-art methods, described below. An R package, called `warpMix`, has been developed for the proposed method and is available at https://cran.r-project.org/web/packages/warpMix/index.html.

4.1 Description of the simulation settings

4.1.1 Warping functions

The warping process is the same in most of the settings (with exception of Model M2), and defined via (4) and (5). Three different interior knots $\{0.2, 0.5, 0.7\}$ are used for a basis of cubic splines for h^{-1}. So in this setting $\kappa_h = 3$, $p_h = 3$ and $r = \kappa_h + p_h + 1 = 7$.

The random variables $(\boldsymbol{\theta}_i)_{i=1,\dots,n}$ are distributed according to $\mathcal{N}_r(\mathbf{0}_r, \Sigma^{\boldsymbol{\theta}})$, with $\Sigma^{\boldsymbol{\theta}} = \Sigma^{\mathbf{E}} + \sigma^2_{\tilde{\boldsymbol{\varepsilon}}}\mathbf{I}$, where $\sigma^2_{\tilde{\boldsymbol{\varepsilon}}} = 10^{-3}$, and $\Sigma^{\mathbf{E}}$ a diagonal matrix with elements $\{2, 0.8, 0.4, 0.3, 0.4, 0.8, 2\}$.

Figure 2 depicts a sample of size 100 of the warping function, and the empirical covariance matrix of $(w^{-1}(.;\boldsymbol{\theta}_i))_{i=1,\dots,n}$ computed on this sample. This highlights the differences between the correlation in $(\boldsymbol{\theta}_i)_{i=1,\dots,n}$ and that in $(w^{-1}(.;\boldsymbol{\theta}_i))_{i=1,\dots,n}$. Note that due to the random warping structure, there is a variability induced on the whole time period.

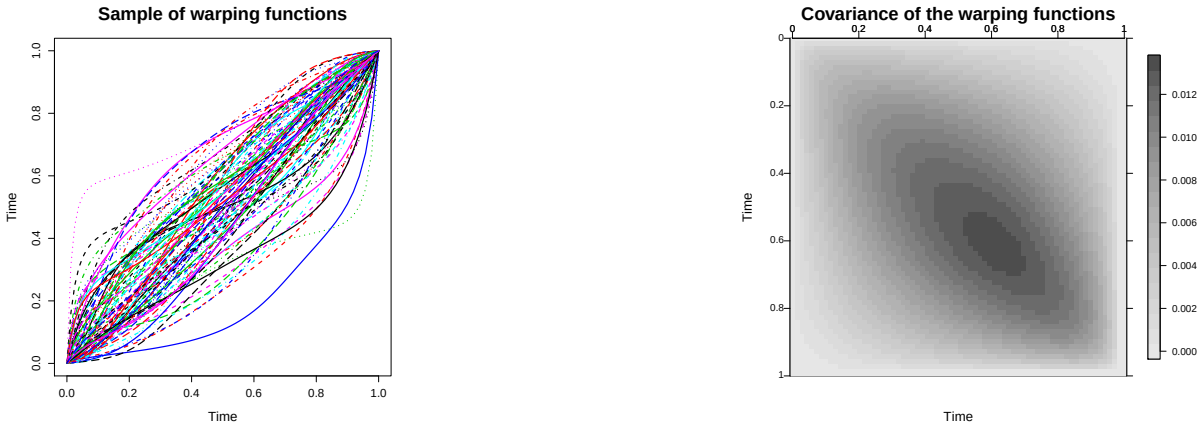

Figure 2: A sample of size 100 of warping functions (left), and the empirical covariances of these functions (right).

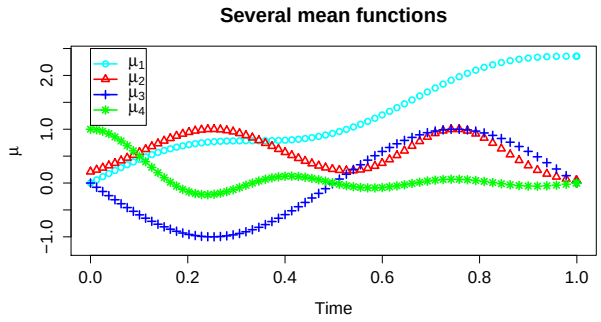

Figure 3: The mean functions μ in the simulation study.

4.1.2 Elements of the functional model

The elements determining the functional model are the function μ and the individual random effects U_i. For the mean function μ we consider four different functions. For $t \in [0,1]$,

$$\mu_1(t) = \{\sin(3\pi t) + 3\pi t\}/4, \qquad \mu_2(t) = \exp^{-(t-0.25)^2/0.04} + \exp^{-(t-0.75)^2/0.02},$$
$$\mu_3(t) = \cos(2\pi t + \pi/2), \qquad \mu_4(t) = \sin(6\pi t)/(6\pi t).$$

These functions are plotted in Figure 3.

The modeling framework in Section 2 assumes that the functions μ and U_i are well approximated using a B-spline basis. This is in practice not always the case, for example when a too limited number of knots is considered in the B-spline bases. In the simulation study we present results on the B-spline approximations of the μ-functions, denoted by $\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3$ and $\tilde{\mu}_4$ respectively. To illustrate the impact of modeling bias, we provide for the fourth function simulation results for its B-spline approximation $\tilde{\mu}_4$ as well as for the function μ_4 itself. We refer to model (2) with mean function $\tilde{\mu}_k$ $(k = 1, 2, 3, 4)$ as model $\widetilde{\mathrm{M}}_k$, and with mean function μ_4 as model M_4.

We consider a low-dimensional setting in which $n = 100$ and $T_i = 70$, as well as a high-dimensional setting in which $n = 200$ and $T_i = 150$. We use $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$, $\widetilde{\mathrm{M}}_2^{\mathrm{HD}}$, $\widetilde{\mathrm{M}}_3^{\mathrm{HD}}$ and $\widetilde{\mathrm{M}}_4^{\mathrm{HD}}$ to refer to the high-dimensional sample setting.

The estimators of μ and U are computed using quadratic splines ($p_\mu = p_{U_i} = 2$), with interior knots at $\{0.12, 0.24, 0.36, 0.48, 0.60, 0.72, 0.84\}$, so that $K_\mu = K_{U_i} = 7$, and $m_\mu = m_{U_i} = 7 + 2 + 1 = 10$. The individual effects U_i in the functional model have a centered multivariate normal distribution with diagonal isotropic covariance matrix $\Sigma^U = 0.1\mathbf{I}_{10}$, except for models $\widetilde{\mathrm{M}}_4$ and $\widetilde{\mathrm{M}}_4^{\mathrm{HD}}$ which are harder to fit, where we use $\Sigma^U = 0.05\mathbf{I}_{10}$. The variance of $\boldsymbol{\varepsilon}$ in the functional linear model equals $\sigma^2_{\boldsymbol{\varepsilon}} = 0.02$.

In the numerical study, we simulate 100 times from each setting, and report the evaluation criteria based on these 100 simulated samples.

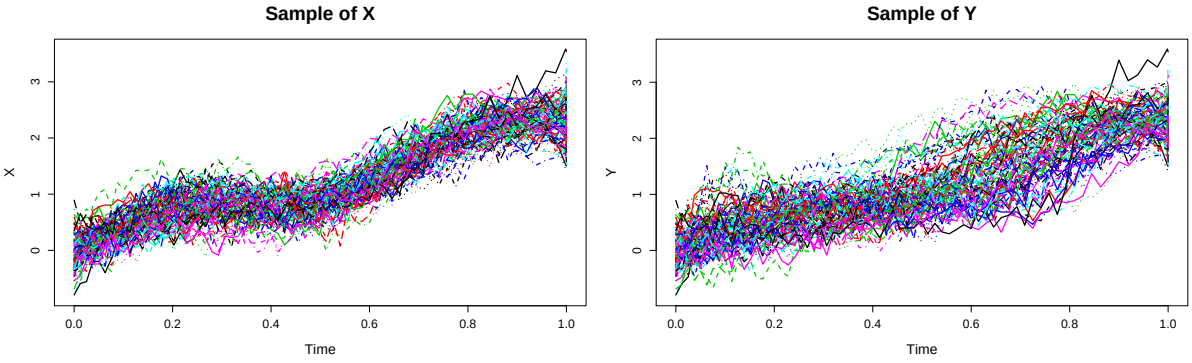

Figure 4: A sample of the warped process X (left) and the un-warped process Y (right) for M_1.

4.1.3 Variability

To generate the data, we first construct a sample of the process $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, defined in (8) and then un-warp them via (9) and the warping function described in Section 4.1.1. To understand the variability induced by each modeling aspect, we plot in Figure 4, a warped sample and the un-warped sample for model M_1.

The signal-to-noise ratio expresses the ratio of the variability caused by the signal $[\mu\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i)\} + U_i\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i)\}]$ and that due to the noise $\varepsilon_{i,j}$

$$\mathrm{SNR}(t_{i,j}) = \frac{\mathrm{Var}\left[\mu\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i)\} + U_i\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i)\}\right]}{\mathrm{Var}(\varepsilon_{i,j})}.$$

To compute the numerator we use the conditional variance formula, for a random variable V seen as a function of two random variables U and $\boldsymbol{\theta}$,

$$\mathrm{Var}_{(\boldsymbol{\theta},U)}(V) = \mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{Var}_U(V|\boldsymbol{\theta})\} + \mathrm{Var}_{\boldsymbol{\theta}}\{E_U(V|\boldsymbol{\theta})\}.$$

For each given time point $t_{i,j}$ we compute this SNR function 50 times to get 50 values for SNR at each time point. To compute the function once, we proceed as follows. For a fixed $\boldsymbol{\theta}$, we compute the empirical conditional variance $\mathrm{Var}_U(Z|\boldsymbol{\theta})$ and the conditional expectation $E_U(Z|\boldsymbol{\theta})$ over a sample of size 100. By varying $\boldsymbol{\theta}$ 60 times, we compute the global variance. This whole process is then repeated 50 times. In Figure 5 the resulting approximations for the SNR functions for models $\widetilde{\mathrm{M}}_1$ and $\widetilde{\mathrm{M}}_2$ are plotted. For higher values of SNR we expect the estimation problem to be somewhat easier. Some caution regarding this interpretation is needed though. In our functional mixed effects model there are several sources of variability in the signal part (the individual effect related to U_i and the warping effect due to $\boldsymbol{\theta}_i$). The SNR-criterion does not distinguish between these variabilities, and just considers the global signal variability against the error variability. Note from the SNR plots in Figure 5 that the estimation task can be harder in some time-regions. At the endpoints of the interval $[0, 1]$ the SNR-values for the different models are equal, since the warping is not effecting these parts, and the only effect is coming from the covariance matrix Σ^U, the noise variance $\sigma^2_{\boldsymbol{\varepsilon}}$, and their relative contribution.

4.2 Comparison with existing methods and performance criteria

To illustrate the numerical performance of the proposed method, we compare with four methods available in the literature.

Since our nonlinear functional mixed effects model is closely related to that of Gervini and Carter (2014) with major differences as indicated in Section 1, we include a comparison with this method. Some procedure parameters have to be chosen in the method of Gervini and Carter (2014): we took $p = q = 1$ for the number of components in the Karhunen-Loève decompositions; $\lambda = 1$ for the regularization parameter; and $\boldsymbol{\tau}_0 = \{0.3, 0.6\}$ as the set of average landmarks. Their estimation method involves a Monte Carlo approximation part, for which we considered 100 iterations; and an EM algorithm part in which we also considered at most 100 iterations. Convergence was said to be reached when the difference in norm between estimated parameters in two consecutive iteration steps was less than 10^{-2}. We also would like to mention that in our simulation study we use a rewritten Matlab version of the original

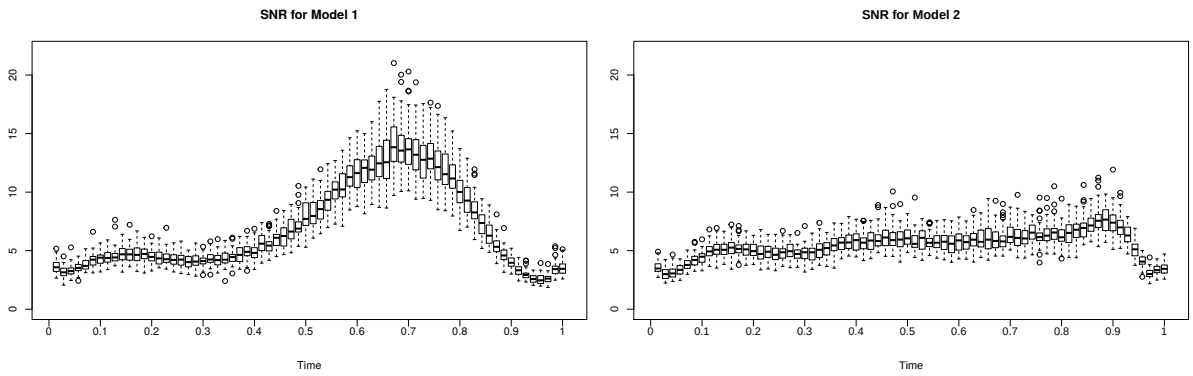

Figure 5: SNR functions for $\widetilde{\mathrm{M}}_1$ (left) and $\widetilde{\mathrm{M}}_2$ (right).

Table 1: Evaluation criteria for the estimation tasks.

quantity of interest	evaluation criterion
μ	$\Delta_\mu = \sum_{j=1}^{T-1}[\hat{\mu}(t_j) - \mu(t_j)]^2(t_{j+1} - t_j)$
w	$\Delta_w = \sum_{j=1}^{T-1}[w^{-1}(t_j; \hat{\boldsymbol{\theta}}_0) - t_j]^2(t_{j+1} - t_j)$
Σ^U	$\Delta_U = \mathrm{Tr}(\hat{\Sigma}^U - \Sigma^U)$
$\Sigma^{\boldsymbol{\theta}}$	$\Delta_{\boldsymbol{\theta}} = \mathrm{Tr}(\hat{\Sigma}^{\boldsymbol{\theta}} - \Sigma^{\boldsymbol{\theta}})$
σ_ε^2	$\Delta_{\boldsymbol{\varepsilon}} = \lvert\hat{\sigma}_{\boldsymbol{\varepsilon}}^2 - \sigma_{\boldsymbol{\varepsilon}}^2\rvert$

Fortran code used in Gervini and Carter (2014), since the latter was no longer running properly. The use of the Matlab code can make computations a bit slower.

The elastic square-root slope is a promising framework, so we include a comparison with the method developed in Tucker et al. (2013). We use the default settings: no elasticity, Karcher mean, do not smooth the data and at most 20 iterations. We use the code available in the R package `fdasrvf`.

Bayesian methods are also of interest, and we choose to compare with Cheng et al. (2016), also available in the R package `fdasrvf`. Also here we considered the default settings: 150000 iterations and a uniform prior distribution.

Finally, we compare the performances with that of the algorithm of Sangalli et al. (2010), available in the R package `fdakma`, that allows for clustering misaligned data. We assume that there is no cluster, consider affine alignment and compute the similarity through the cosine of the angle between the two function.

Since the available inference in those studies does not fully match our modeling inference, we can only report on the comparison related to estimating μ.

To evaluate the estimation performance for the various components of the target $(\boldsymbol{\alpha}^\mu, \sigma_{\boldsymbol{\varepsilon}}^2, \underset{\sim}{\Sigma}^U, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ we need some criteria. Note that the modeling framework involves two unknown functions, namely the overall mean function μ and the warping function w, unknown matrices Σ^U, $\Sigma^{\boldsymbol{\theta}}$, as well as the unknown variance σ_ε^2. For each sample we obtain estimates $\hat{\mu}$, $\hat{w}$, $\hat{\Sigma}^U$, $\hat{\Sigma}^{\boldsymbol{\theta}}$ and $\hat{\sigma}_\varepsilon^2$. Since in our simulation setting we have the same observational time points for each individual curve, i.e. $t_{i,j} = t_j$, and $j = 1, \ldots, T_i$, with $T_i = T$, we use the criteria in Table 1 to evaluate the estimation performance in each sample. Herein $\mathrm{Tr}(\mathbf{A})$ denotes the trace of a matrix $\mathbf{A}$.

For each simulated sample we calculate the estimates, and the corresponding evaluation criteria of Table 1. To report on the bias of an estimator, we compute the empirical mean of a criterion over the 100 simulations. To report on the variance of an estimator, we proceed as follows. For example, when estimating the function μ, we calculate in each point t_j the empirical mean over all 100 estimated values of $\mu(t_j)$ and denoting this by $\bar{\mu}(t_j)$. For each simulated sample we then calculate $\bar{\Delta}_\mu = \sum_{j=1}^{T-1}[\hat{\mu}(t_j) - \bar{\mu}(t_j)]^2(t_{j+1} - t_j)$. The empirical variance of the estimator for μ is then computed by taking the average over the 100 obtained $\bar{\Delta}_\mu$ values. In a similar way we obtain $\bar{\Delta}_w$, $\bar{\Delta}_U$, $\bar{\Delta}_{\boldsymbol{\theta}}$ and $\bar{\Delta}_{\boldsymbol{\varepsilon}}$.

A final remark is that for $\Delta_{\boldsymbol{\theta}}$ and $\bar{\Delta}_{\boldsymbol{\theta}}$, we use medians rather than means across all simulations as a measure of central position, since sometimes estimation of some components of $\Sigma^{\boldsymbol{\theta}}$ resulted in large outlying values. However, even in the latter cases the quality of the estimated warping function was still

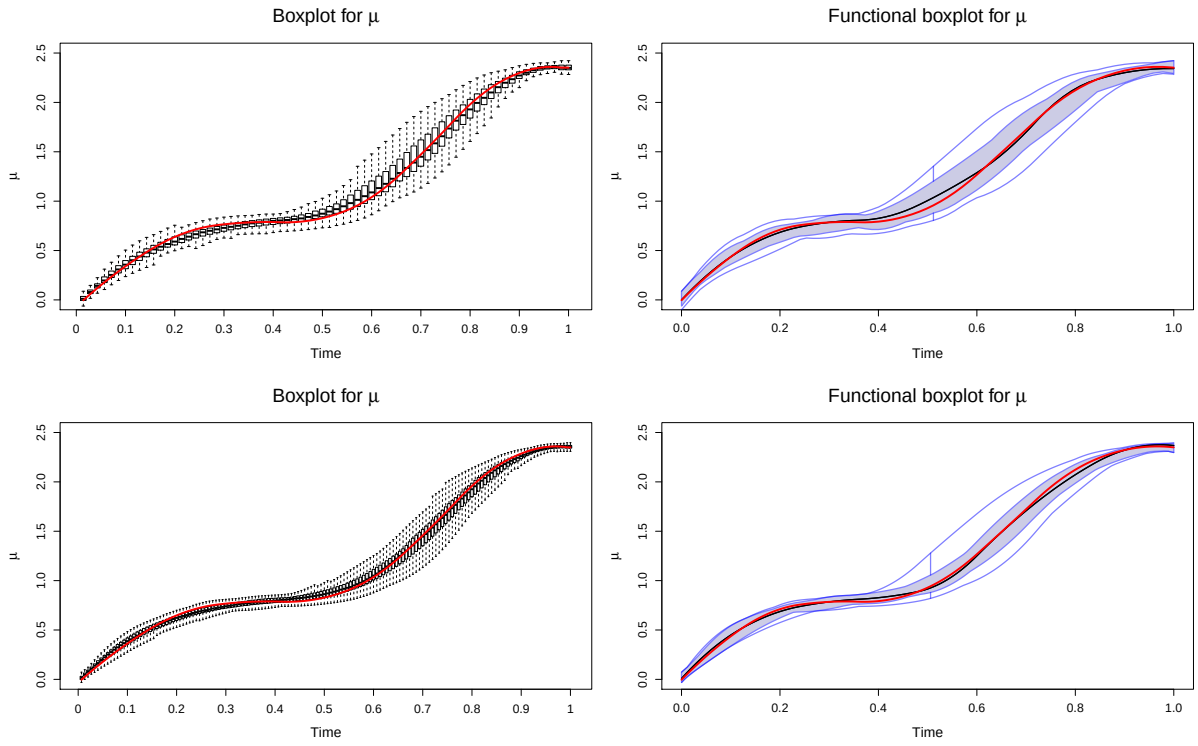

Figure 6: Comparison between $\tilde{\mu}_1$ and $\hat{\mu}_1$ for model $\widetilde{\mathrm{M}}_1$ ($n = 100$ and $T = 70$, top) and for model $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$ ($n = 200$ and $T = 150$, bottom) using the proposed method.

Table 2: Simulation results for the proposed procedure for models $\widetilde{\mathrm{M}}_1$ and $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$.

criterion	$\widetilde{\mathrm{M}}_1$		$\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$	
	bias	variance	bias	variance
Δ_μ	0.0102	0.0090	0.0055	0.0065
Δ_ε	0.0089	0.0003	0.0074	0.0004
Δ_U	0.2349	0.1453	0.1880	0.1070
$\Delta_{\boldsymbol{\theta}}$	6.5050	779.6355	4.7537	138.9394
Δ_w	0.0020	0.0019	0.0007	0.0009

very good, as will be seen from the reported results.

4.3 Simulation results for the proposed method

Models $\widetilde{\mathrm{M}}_1$ and $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$. Figure 6 depicts the simulation results for estimating $\tilde{\mu}_1$ in $\widetilde{\mathrm{M}}_1$ and $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$. In the left panels we depict, for each time point t_j, the boxplots of the obtained estimated values for $\tilde{\mu}_1(t_j)$, whereas in the right panels we use a functional boxplot, as developed in Sun and Genton (2011). The true curve $\tilde{\mu}_1$ is in all plots presented as the solid (red) curve. The black solid curve in the centre of the functional boxplots indicates the deepest function among all estimated mean functions.

The quality of estimating $\tilde{\mu}_1$ is quite good for the proposed method. Passing from low dimension to high dimension (from the top row to the bottom row plots), we see that the results improve for larger values of n and T. Note that the largest variability occurs in the region where there was also most variability noticed in the SNR plot for model $\widetilde{\mathrm{M}}_1$ in Figure 5. Table 2 further summarizes the simulation results for models $\widetilde{\mathrm{M}}_1$ and $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$. The results on estimation of μ are in correspondence with what was observed from Figure 6. Note that in estimation of $\Sigma^{\boldsymbol{\theta}}$ there are quite some extreme estimation results. However, the resulting estimation of the warping function w is still good, as can be noticed from the last row in Table 2.

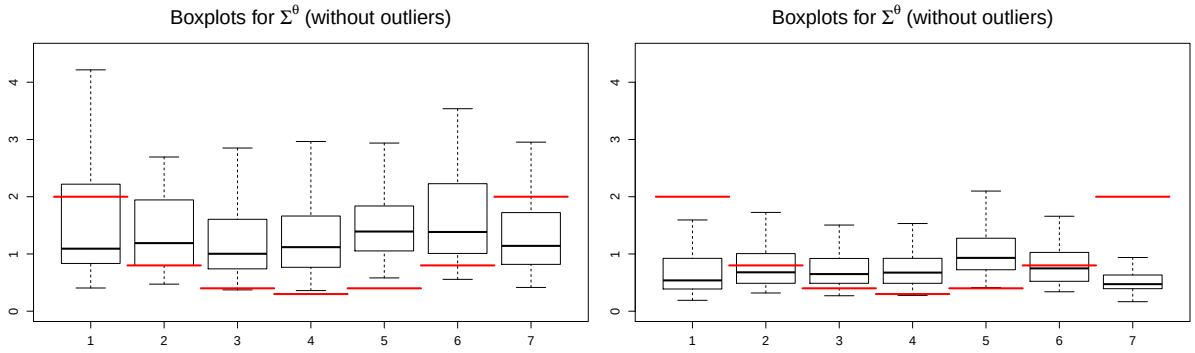

Figure 7: Simulation results for the components of $(\hat{\Sigma}^{\boldsymbol{\theta}}_{i,i})_{1\leq i\leq 7}$ for the proposed procedure for model $\widetilde{\mathrm{M}}_1$ (left) and $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$ (right). Boxplot without the outliers. True coefficient values: red horizontal lines.

Table 3: Simulation results for $\Sigma^{\boldsymbol{\theta}}$ using the proposed procedure for models $\widetilde{\mathrm{M}}_1$ and $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$.

Model	maximum of estimated diagonal components						
$\widetilde{\mathrm{M}}_1$	30976.0730	6141.906	953.239	333.466	325.040	345.714	290.960
$\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$	4664.249	193.563	4.735	5.023	46.427	5.329	1.937

In Figure 7 we present boxplots of the estimation results for the components of $\Sigma^{\boldsymbol{\theta}}$ for models $\widetilde{\mathrm{M}}_1$ (left) and $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$ (right), with the true component values indicated as red horizontal lines. Outliers have been excluded for plotting the boxplots for clarity of presentation. To complement these boxplots, we summarize in Table 3 the maximum (across simulations) of the estimated values for each of the seven components of $\Sigma^{\boldsymbol{\theta}}$. Note that the most extreme values occur for the first coefficient. For larger n and T there are less extreme estimates.

Next, we focus on estimating the individual curve amplitude variability, which is captured by the estimation of the ten diagonal components of Σ^U. In Figure 8 we provide boxplots of the estimation results. The horizontal red line presents the true value 0.10 for all diagonal components. As can be seen the estimation results tend to be better for larger value of n and T, as expected.

Models $\widetilde{\mathbf{M}}_2^{\mathbf{w_{CDG}}}$ and $\widetilde{\mathbf{M}}_2^{\mathbf{w_{GC}}}$. To study the finite-sample performance of the proposed estimation method when there is a misspecification with respect to the warping function, we consider the model with $\tilde{\mu}_2$ and simulate data under two different warping schemes:

- scheme $\mathrm{w}_{\mathrm{CDG}}$: the warping scheme of Section 2.3;
- scheme w_{GC}: the warping scheme of Gervini and Carter (2014).

In the scheme w_{GC}, $\boldsymbol{\theta}_i$ are generated via a linear mixed effects model, and then mapped into the set of landmarks using a Jupp transform. This is followed by interpolation by cubic splines to get to the corresponding parameters. Simulations were carried out from the two models, referred to as models $\widetilde{\mathrm{M}}_2^{\mathrm{w_{CDG}}}$ and $\widetilde{\mathrm{M}}_2^{\mathrm{w_{GC}}}$.

Table 4 summarizes the simulation results for all elements in the functional mixed effects model. Overall conclusions remain as above. Note that also under the misspecified warping scheme w_{GC} the proposed method continues to perform very well.

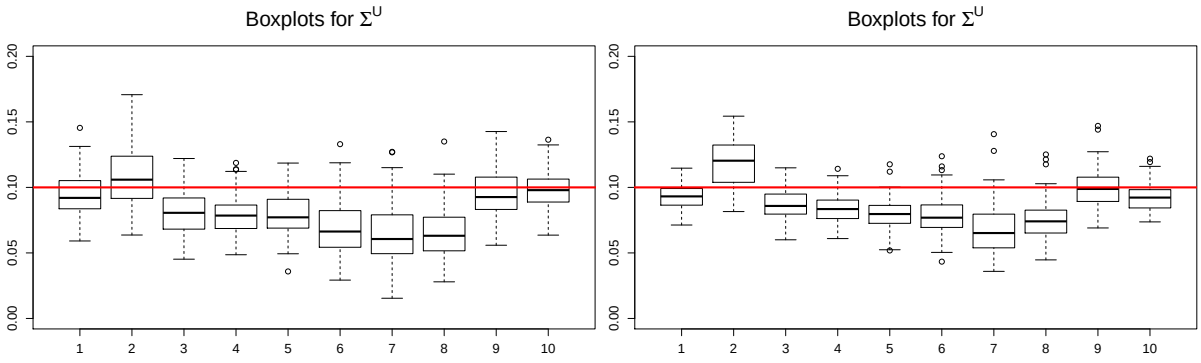

Figure 8: Simulation results for the components of $(\hat{\Sigma}^U_{i,i})_{1\leq i\leq 10}$ for models $\widetilde{\mathrm{M}}_1$ (left) and $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$ (right) for the proposed procedure. The horizontal red line present the true value.

Table 4: Simulation results for the proposed procedure for models $\widetilde{\mathrm{M}}_2^{\mathrm{wCDG}}$, $\widetilde{\mathrm{M}}_2^{\mathrm{wGC}}$

criterion	$\widetilde{\mathrm{M}}_2^{\mathrm{wCDG}}$		$\widetilde{\mathrm{M}}_2^{\mathrm{wGC}}$	
	bias	variance	bias	variance
Δ_μ	0.0230	0.0136	0.0053	0.0052
Δ_ε	0.0079	0.0009	0.0089	0.0003
Δ_U	0.4044	0.2809	0.2552	0.1388
$\Delta_{\boldsymbol{\theta}}$	16.3997	497.9732	3.7417	42.9801
Δ_w	0.0012	0.0027	0.0003	0.0003

Models $\widetilde{\mathrm{M}}_3$ and $\widetilde{\mathrm{M}}_3^{\mathrm{HD}}$, and Models M_4, $\widetilde{\mathrm{M}}_4$ and $\widetilde{\mathrm{M}}_3^{\mathrm{HD}}$. Table 5 presents the simulation results for the low- and high-dimensional settings for model $\widetilde{\mathrm{M}}_3$. Also in these settings the method performs well.

Table 5: Simulation results for the proposed procedure for models $\widetilde{\mathrm{M}}_3$ and $\widetilde{\mathrm{M}}_3^{\mathrm{HD}}$

criterion	$\widetilde{\mathrm{M}}_3$		$\widetilde{\mathrm{M}}_3^{\mathrm{HD}}$	
	bias	variance	bias	variance
Δ_μ	0.0183	0.0153	0.0126	0.0105
Δ_ε	0.0087	0.0003	0.0070	0.0005
Δ_U	0.2554	0.2111	0.2400	0.1622
$\Delta_{\boldsymbol{\theta}}$	6.1263	1400.3432	4.8580	865.6914
Δ_w	0.0021	0.0020	0.0013	0.0018

For the fourth model we include simulation results (in the low-dimensional sample setting) when simulating from the unprojected function μ_4, for which the B-spline approximation induces a modeling bias. As can be seen from columns 2—5 in Table 6 there is only a little loss in performance when modeling bias is present.

Table 6: Simulation results for the proposed procedure for models M_4, $\widetilde{\mathrm{M}}_4$ and $\widetilde{\mathrm{M}}_4^{\mathrm{HD}}$.

criterion	M_4		$\widetilde{\mathrm{M}}_4$		$\widetilde{\mathrm{M}}_4^{\mathrm{HD}}$	
	bias	variance	bias	variance	bias	variance
Δ_μ	0.0110	0.0070	0.0088	0.0044	0.0057	0.0018
Δ_ε	0.0138	0.0003	0.0085	0.0004	0.0075	0.0003
Δ_U	0.4398	0.1033	0.3906	0.1241	0.3999	0.0942
$\Delta_{\boldsymbol{\theta}}$	8.2414	2139.4377	4.0467	602.5896	4.3262	299.9634
Δ_w	0.0043	0.0032	0.0022	0.0012	0.0011	0.0005

4.4 Comparison with available methods.

We compare the four methods introduced in Section 4.2 with the proposed one on models $\widetilde{\mathrm{M}}_1$, $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$, $\widetilde{\mathrm{M}}_2^{\mathrm{wCDG}}$ and $\widetilde{\mathrm{M}}_2^{\mathrm{wGC}}$. The simulation results are summarized in Table 7. Note that the proposed method often has low/lowest bias, but at the price of having a larger estimation variance. On model $\widetilde{\mathrm{M}}_1$, the method `fdakma` performs the best, with a very low variance, but it has a comparable performance (in terms of bias) to the proposed method for $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$. On models $\widetilde{\mathrm{M}}_2^{\mathrm{wCDG}}$ and $\widetilde{\mathrm{M}}_2^{\mathrm{wGC}}$, our method has particularly good results in mean, but with a larger variance.

Table 7: Simulation results for $\mu(\cdot)$ for the proposed and competitive methods. Method (3)=Bayesian warping; Method (4)= elastic square-root slope.

Simulation results for models $\widetilde{\mathrm{M}}_1$, $\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$, $\widetilde{\mathrm{M}}_2^{\mathrm{wCDG}}$ and $\widetilde{\mathrm{M}}_2^{\mathrm{wGC}}$

Method	$\widetilde{\mathrm{M}}_1$		$\widetilde{\mathrm{M}}_1^{\mathrm{HD}}$		$\widetilde{\mathrm{M}}_2^{\mathrm{wCDG}}$		$\widetilde{\mathrm{M}}_2^{\mathrm{wGC}}$	
	bias	variance	bias	variance	bias	variance	bias	variance
proposed	0.0102	0.0090	0.0055	0.0065	0.0230	0.0236	0.0053	0.0052
GC	0.0454	0.0043	0.0588	0.0048	0.0703	0.0134	0.0280	0.0014
(3)	0.0435	0.0009	0.0201	0.0001	0.0493	0.0003	0.0423	0.0001
(4)	0.1062	0.0011	0.0745	0.0009	0.1277	0.0014	0.1245	0.0011
fdakma	0.0075	7.10^{-6}	0.0069	3.10^{-6}	0.0281	2.10^{-5}	0.0283	1.10^{-5}

Simulation results for models $\widetilde{\mathrm{M}}_3$, $\widetilde{\mathrm{M}}_3^{\mathrm{HD}}$, $\widetilde{\mathrm{M}}_4$ and $\widetilde{\mathrm{M}}_4^{\mathrm{HD}}$

Method	$\widetilde{\mathrm{M}}_3$		$\widetilde{\mathrm{M}}_3^{\mathrm{HD}}$		$\widetilde{\mathrm{M}}_4$		$\widetilde{\mathrm{M}}_4^{\mathrm{HD}}$	
	bias	variance	bias	variance	bias	variance	bias	variance
proposed	0.0183	0.0153	0.0126	0.0105	0.0088	0.0044	0.0057	0.0018
(3)	0.0794	0.0024	0.0491	0.0002	0.0280	0.0003	0.0212	0.0001
(4)	0.1851	0.0009	0.1951	0.0007	0.1555	0.0016	0.1754	0.0026
fdakma	0.0277	4.10^{-5}	0.0272	2.10^{-5}	0.0113	5.10^{-6}	0.0104	2.10^{-6}

As GC's method is very slow (i.e high computational cost) and does not provide very good results whereas the modelling is close to the proposed one, we restrict further comparisons, for Models $\widetilde{\mathrm{M}}_3$, $\widetilde{\mathrm{M}}_3^{\mathrm{HD}}$, $\widetilde{\mathrm{M}}_4$ and $\widetilde{\mathrm{M}}_4^{\mathrm{HD}}$, to the other three methods. On all models $\widetilde{\mathrm{M}}_3$, $\widetilde{\mathrm{M}}_4^{\mathrm{HD}}$, $\widetilde{\mathrm{M}}_4$ and $\widetilde{\mathrm{M}}_4^{\mathrm{HD}}$, fdakma performs the best among the competitive methods, followed by the Bayesian warping method (method (3)), but both are less good than the proposed method in terms of bias. Finally, we see that the elastic square-root slope method (method (4) in the table) does not perform well on those simulated datasets.

5 Real data analysis

We analyze the Pinch Force dataset, available in the R package `fda`. These data were described and analyzed in Ramsay et al. (1995). The data consist of measurements, at every second millisecond, on the exerted force (in Newton) during a period of 0.3 seconds. The resulting measurements consist of 20 curves recorded on 151 points in $[0, 0.3]$. See Figure 1. For convenience the data were rescaled to the domain $[0, 1]$.

We analyzed these data, using B-splines of degree 2 for μ and U (i.e. $p_\mu = p_{U_i} = 2$), with interior knots $\{0.25, 0.50, 0.75\}$, resulting in six B-spline basis functions. For the function determining the warping in (5) we use B-splines of degree 3 (i.e. $p_h = 3$) and the same set of interior knots $\{0.25, 0.50, 0.75\}$, leading to seven B-spline basis functions for h^{-1}.

From the analysis with the proposed method, we get the estimated individual warping functions as in the left panel of Figure 9, and the warped (aligned) functions $X_{i,j}$ for each individual (right panel). The estimated covariance matrix $\hat{\Sigma}^U$ (respectively $\hat{\Sigma}^{\boldsymbol{\theta}}$) is presented in the left (respectively right) panel of Figure 10. From this, we observe that there is more time variability induced by the coefficient of the second B-spline basis function in the decomposition of $h^{-1}(., \boldsymbol{\theta})$, whereas there is more amplitude variability caused by the coefficient associated to the third basis function in the decomposition of U, see also Figure 9 (right panel).

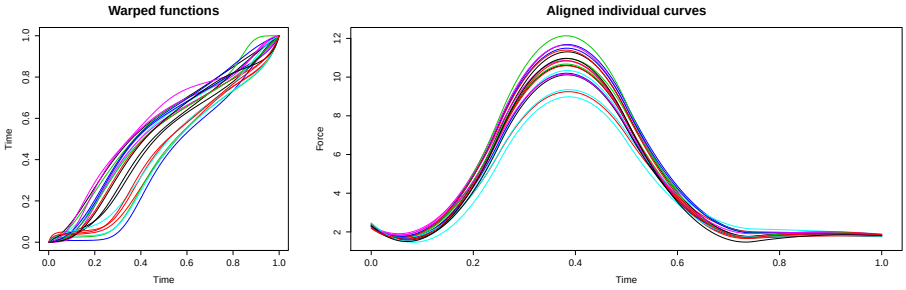

Figure 9: Estimated warping functions (left) and warped individual curves (right).

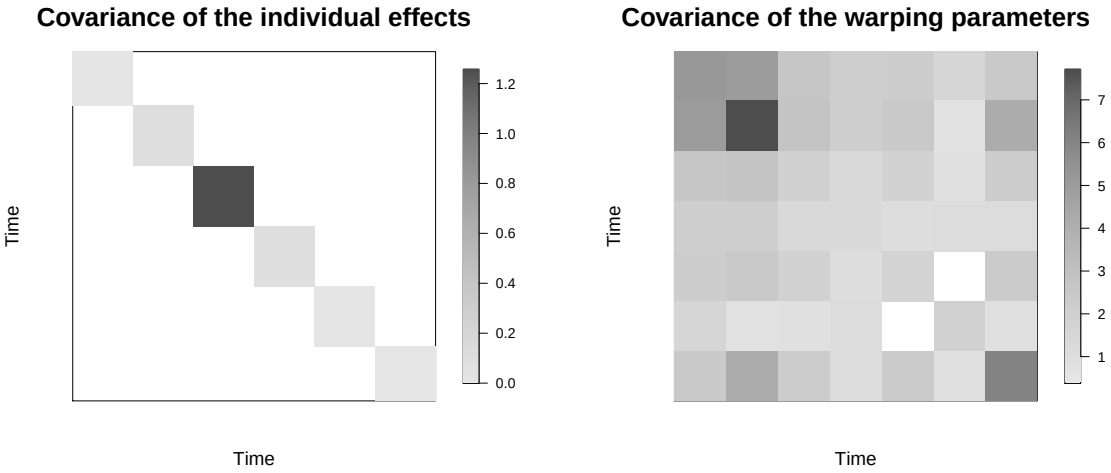

Figure 10: Estimated covariance matrices $\hat{\Sigma}^U$ (left) and absolute value of $\hat{\Sigma}^{\boldsymbol{\theta}}$ (right).

6 Conclusion and further discussion

In this paper we considered a nonlinear mixed effect model for functional data. We apply a B-splines approximation on three different levels: on the inverse of the warping function describing the individual phase variability; on the global mean function and on the individual amplitude random effects. Random effects enter to model the individual amplitude as well as the phase variability. The main advantage of the proposed method is that it avoids the (costly) choice of landmarks, and that we can provide important theoretical support for the procedure: (i) convergence of the iterative algorithm to the target function(s); (ii) consistency and asymptotic normality of the estimators.

In this paper we considered the discrete T_i time points to be fixed (non-random). However the methodology could be generalized fairly easily to random time points. Typically one would then need to assume that the distribution of the random time points is regular enough (meaning that there are no empty regions in the observed pattern of discretized time points). This would require, for example, an adaptation of the criteria used in Sections 3.2 and 3.3.2. An analysis of the assumptions, particularly modeling assumptions of the noise, with the aim to see the robustness of the method, would be of interest.We postpone this analysis to an experimental work.

A Proof of results in Section 2

A.1 Proof of Lemma 1

Let $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ such that $t = w(w^{-1}(t;\boldsymbol{\theta}^1);\boldsymbol{\theta}^2)$. Then it follows that,

$$\begin{aligned} w^{-1}(t;\boldsymbol{\theta}^2) = \frac{\int_0^t \exp(h^{-1}(u;\boldsymbol{\theta}^2))du}{\int_0^1 \exp(h^{-1}(u;\boldsymbol{\theta}^2))du} &= \frac{\int_0^t \exp(h^{-1}(u;\boldsymbol{\theta}^1))du}{\int_0^1 \exp(h^{-1}(u;\boldsymbol{\theta}^1))du} = w^{-1}(t;\boldsymbol{\theta}^1) \\ \Leftrightarrow \qquad \frac{\int_0^t \exp(h^{-1}(u;\boldsymbol{\theta}^2))du}{\int_0^t \exp(h^{-1}(u;\boldsymbol{\theta}^1))du} &= \frac{\int_0^1 \exp(h^{-1}(u;\boldsymbol{\theta}^2))du}{\int_0^1 \exp(h^{-1}(u;\boldsymbol{\theta}^1))du} \stackrel{\text{def}}{=} \delta(\boldsymbol{\theta}^2,\boldsymbol{\theta}^1) \\ \Leftrightarrow \qquad \int_0^t \left[\delta(\boldsymbol{\theta}^2,\boldsymbol{\theta}^1)\exp(h^{-1}(u;\boldsymbol{\theta}^1)) - \exp(h^{-1}(u;\boldsymbol{\theta}^2))\right] du &= 0. \end{aligned}$$

As this equation is true whatever the value of $t \in [0,1]$, we have that the integrand is equal to 0 for every $u \in [0,1]$ except for possibly a countable number of points. As B-splines are continuous, it is equal to 0 for every $u \in [0,1]$. It holds that,

$$\log(\delta(\boldsymbol{\theta}^2,\boldsymbol{\theta}^1)) = h^{-1}(u;\boldsymbol{\theta}^2) - h^{-1}(u;\boldsymbol{\theta}^1) = \sum_{l=-p_h}^{K_h} (\theta_l^2 - \theta_l^1)\bar{B}_l^h(u;\boldsymbol{\kappa}^h).$$

As the left hand side does not depend on u, so the right hand side should be equal to 0 for all u. As $(B_l^{\tilde{h}})_l$ is a B-spline basis, it induces that $\boldsymbol{\theta}^2 = \boldsymbol{\theta}^1$. □

A.2 Proof of Theorem 1

First, suppose that $\mathbf{H}_i^U = \mathbf{I}_{T_i}$. Let $\Sigma_{\boldsymbol{\varepsilon}} = \sigma_{\boldsymbol{\varepsilon}}^2\mathbf{I}_{T_i}$, and choose $0 < \tilde{\sigma}_{\boldsymbol{\varepsilon}}^2 < \sigma_{\boldsymbol{\varepsilon}}^2$. Then, $\mathbf{H}_i^U(\Sigma_{\boldsymbol{\varepsilon}} - \tilde{\Sigma}_{\boldsymbol{\varepsilon}}) = \Sigma_{\boldsymbol{\varepsilon}} - \tilde{\Sigma}_{\boldsymbol{\varepsilon}}$. Since $\tilde{\sigma}_{\boldsymbol{\varepsilon}}^2 - \sigma_{\boldsymbol{\varepsilon}}^2 < 0$, and $(\mathbf{B}_i^U)^\top\mathbf{B}_i^U \neq \mathbf{0}_{m_{U_i}}$,

$$\tilde{\Sigma}^U = \Sigma^U + (\sigma_{\boldsymbol{\varepsilon}}^2 - \tilde{\sigma}_{\boldsymbol{\varepsilon}}^2)\{(\mathbf{B}_i^U)^\top\mathbf{B}_i^U\}^{-1}$$

is semi positive, definite, and is diagonal. Thus, we have found two parameters $(\Sigma_{\boldsymbol{\varepsilon}},\Sigma^U)$ and $(\tilde{\Sigma}_{\boldsymbol{\varepsilon}},\tilde{\Sigma}^U)$ which define the same model. Hence we do not have identifiability.

Suppose now that model (8) is not identifiable. Then, according to Theorem 4.1 in Wang (2013), for all $(\Sigma_{\boldsymbol{\varepsilon}},\Sigma^U)$, there exists $(\tilde{\Sigma}_{\boldsymbol{\varepsilon}},\tilde{\Sigma}^U) \neq (\Sigma_{\boldsymbol{\varepsilon}},\Sigma^U)$ such that

- $(\mathbf{B}_i^U)^\top\Sigma_{\boldsymbol{\varepsilon}}\mathbf{B}_i^U \neq (\mathbf{B}_i^U)^\top\tilde{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{B}_i^U$;
- $\mathbf{H}_i^U(\Sigma_{\boldsymbol{\varepsilon}} - \tilde{\Sigma}_{\boldsymbol{\varepsilon}}) = \Sigma_{\varepsilon} - \tilde{\Sigma}_{\boldsymbol{\varepsilon}}$;
- $\tilde{\Sigma}^U = \Sigma^U + \{(\mathbf{B}_i^U)^\top\mathbf{B}_i^U\}^{-1}(\Sigma_{\boldsymbol{\varepsilon}} - \tilde{\Sigma}_{\boldsymbol{\varepsilon}})\mathbf{B}_i^U\{(\mathbf{B}_i^U)^\top\mathbf{B}_i^U\}^{-1} = \Sigma^U + (\sigma_{\boldsymbol{\varepsilon}}^2 - \tilde{\sigma}_{\boldsymbol{\varepsilon}}^2)\{(\mathbf{B}_i^U)^\top\mathbf{B}_i^U\}^{-1}$.

As $\Sigma_{\boldsymbol{\varepsilon}} \neq \tilde{\Sigma}_{\boldsymbol{\varepsilon}}$, $\mathbf{H}_i^U = \mathbf{I}_{T_i}$.

Since $\tilde{\Sigma}^U \neq \Sigma^U$, one gets $(\mathbf{B}_i^U)^\top(\Sigma_{\boldsymbol{\varepsilon}} - \tilde{\Sigma}_{\boldsymbol{\varepsilon}})\mathbf{B}_i^U = (\sigma_{\boldsymbol{\varepsilon}}^2 - \tilde{\sigma}_{\boldsymbol{\varepsilon}}^2)(\mathbf{B}_i^U)^\top\mathbf{B}_i^U \neq \mathbf{0}_{m_{U_i}}$, which implies $(\mathbf{B}_i^U)^\top\mathbf{B}_i^U \neq \mathbf{0}_{m_{U_i}}$.

Moreover, as $\tilde{\Sigma}^U$ is diagonal, $(\mathbf{B}_i^U)^\top\mathbf{B}_i^U$ must be diagonal. □

A.3 Proof of Theorem 2

We have

$$X_i(t) = \tilde{X}_i\left[w^{-1}\left\{w(t;\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\right\}\right] = \mu(t) + U_i(t) + \varepsilon_i(t).$$

As $X_i(t)$ is Gaussian distributed, the identifiability of the warping process is equivalent to the identifiability of the mean and of the variance of X_i. We thus investigate the expectation and the variance of $X_i(t)$.

- Expectation:

$$\begin{aligned} \mathrm{E}[X_i(t)] &= \mathrm{E}\left[\tilde{X}_i\{w^{-1}\{w(t;\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\}\}\right] \\ &= \mathrm{E}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left[\mathrm{E}[\tilde{X}_i\{w^{-1}\{w(t;\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\}\}|\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i]\right]; \\ \mu(t) &= \mathrm{E}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left[\mu\{w^{-1}\{w(t;\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\}\}\right]; \end{aligned}$$

which is equivalent to, by projecting onto the B-spline basis, for all $l = -p_\mu, \dots, K_\mu$, and for all $j = 1, \dots, T_i$,

$$B^\mu_{l,p_\mu+1}(t_{i,j}; \boldsymbol{\kappa}^\mu) = \mathrm{E}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left[B^\mu_{l,p_\mu+1}\{w^{-1}\{w(t_{i,j};\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\};\boldsymbol{\kappa}^\mu\}\right]$$

or, in a matrix representation, $\mathbf{B}^\mu_i = \mathrm{E}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left\{(\mathbf{B}^\mu_i)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\right\}$.

- Variance.
  By the formula of the conditional variance,

$$\begin{aligned}\mathrm{Var}\left[\tilde{X}_i\{w^{-1}\{w(t;\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\}\}\right] =& \mathrm{E}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left[\mathrm{Var}\left[\tilde{X}_i\{w^{-1}\{w(t;\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\}\}\middle|\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i\right]\right]\\ &+ \mathrm{Var}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left[\mathrm{E}\left[\tilde{X}_i\{w^{-1}\{w(t;\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\}\}\middle|\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i\right]\right]. \qquad (13)\end{aligned}$$

  Now,

$$\begin{aligned}&\mathrm{Var}\left[\left[\tilde{X}_i\{w^{-1}\{w(t_{i,j};\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\}\}\right]_{j=1,\dots,T_i}\right] = \mathrm{E}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left[\{(\mathbf{B}^U_i)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\}^\top\Sigma^{U_i}(\mathbf{B}^U_i)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\right]\\ &+\mathrm{E}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left[\mathrm{Var}\left[\left(\varepsilon_i\{w^{-1}\{w(t_{i,j};\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\}\}\right)_{j=1,\dots,T_i}\right]\right] + \mathrm{Var}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left\{(\mathbf{B}^\mu_i)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\boldsymbol{\alpha}^\mu\right\},\end{aligned}$$

  and

$$\mathrm{Var}\left[\{X_i(t_{i,j})\}_{j=1,\dots,T_i}\right] = (\mathbf{B}^U_i)^\top\Sigma^{U_i}\mathbf{B}^U_i + \mathrm{Var}[\{\varepsilon_i(t_{i,j})\}_{j=1,\dots,T_i}].$$

  As $\mathrm{Var}[\{\varepsilon_i(t_{i,j})\}_{j=1,\dots,T_i}] = \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_{T_i}$, and

$$\mathrm{Var}\left[\left(\varepsilon_i\{w^{-1}\{w(t_{i,j};\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\}\}\right)_{j=1,\dots,T_i}\right] = \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_{T_i},$$

  (13) becomes

$$(\mathbf{B}^U_i)^\top\Sigma^{U_i}\mathbf{B}^U_i = \mathrm{Var}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left\{(\mathbf{B}^\mu_i)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\boldsymbol{\alpha}^\mu\right\} + \mathrm{E}_{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\left[\{(\mathbf{B}^U_i)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\}^\top\Sigma^{U_i}(\mathbf{B}^U_i)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}\right].$$

□

B Proof of results in Section 3

B.1 Proof of Theorem 3

Before passing to the proof of Theorem 3 we study the weighted tail cross products, that are appearing in this context. This is done in Lemmas 2 and 3.

Lemma 2. *The weighted tail cross product of $\mu(w^{-1}(t_{i,j};.)) + U_i(w^{-1}(t_{i,j};.))$ with itself (with respect to T_i) exists and*

$$\begin{aligned}Q_{\boldsymbol{\theta}_i}(\tilde{\boldsymbol{\theta}}_i) =& |\mu(w^{-1}(t_{i,j};\boldsymbol{\theta}_i)) + U_i(w^{-1}(t_{i,j};\boldsymbol{\theta}_i))\\ &- \mu(w^{-1}(t_{i,j};\tilde{\boldsymbol{\theta}}_i)) - U_i(w^{-1}(t_{i,j};\tilde{\boldsymbol{\theta}}_i))|^2\end{aligned}$$

has a unique minimum at $\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i$.

Proof. We assume that μ and U_i can be decomposed onto a spline basis. As every function is defined on $[0,1]$, $\mu \circ w^{-1}$ and $U_i \circ w^{-1}$ are L^2. Then the tail cross product of $\mu(w^{-1}(t_{i,j};\boldsymbol{\theta}_i)) + U_i(w^{-1}(t_{i,j};\boldsymbol{\theta}_i))$ with $\mu(w^{-1}(t_{i,j};\tilde{\boldsymbol{\theta}}_i)) + U_i(w^{-1}(t_{i,j};\tilde{\boldsymbol{\theta}}_i))$ exists because it converges uniformly for $\boldsymbol{\theta}_i \in \mathbb{R}^r$ and $\tilde{\boldsymbol{\theta}}_i \in \mathbb{R}^r$.

The function Q has a unique minimum at $\boldsymbol{\theta}_i$: if $Q(\tilde{\boldsymbol{\theta}}_i) = 0$, for every $t \in [0,1]$,

$$(\mu + U_i)(w^{-1}(t;\tilde{\boldsymbol{\theta}}_i)) = (\mu + U_i)(w^{-1}(t;\boldsymbol{\theta}_i)) \quad \Rightarrow \quad t = w(w^{-1}(t;\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i).$$

Then, as the warping function is injective (see Lemma 1), $\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i$. □

Lemma 3. *For $l = 1, \dots, r$ and $l' = 1, \dots, r$, the derivatives*

$$\frac{\partial \left[\mu(w^{-1}(t_{i,j}; \boldsymbol{\theta}_i)) + U_i(w^{-1}(t_{i,j}; .))\right]}{\partial [\boldsymbol{\theta}_i]_l} \text{ and } \frac{\partial^2 \left[\mu(w^{-1}(t_{i,j}; \boldsymbol{\theta}_i)) + U_i(w^{-1}(t_{i,j}; .))\right]}{\partial [\boldsymbol{\theta}_i]_l \partial [\boldsymbol{\theta}_i]_{l'}}$$

exist and are continuous on $\mathbb{R}^r$ and all weighted tail cross products in between

$$\mu(w^{-1}(t_{i,j}; \boldsymbol{\theta}_i)) + U_i(w^{-1}(t_{i,j}; .))$$

and its first and second derivatives exist.

Proof. Each function is defined on $[0, 1]$, and is either decomposed onto a spline basis, or composed with the exponential function, then there is no problem to intervert derivation and integrals. As explained before, the weighted tail cross products exist because those functions are $L^2([0, 1])$. □

Equipped with Lemmas 2 and 3 we can adapt results from Jennrich (1969) to prove Theorem 3.

Proof of Theorem 3. We consider the following compact set:

$$\mathcal{E}_\alpha(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}) = \left\{x \in \mathbb{R}^r \mid (\boldsymbol{\theta}_0 - x)^t (\Sigma^{\boldsymbol{\theta}})^{-1} (\boldsymbol{\theta}_0 - x) \leq \chi^2_r(1 - \alpha)\right\}$$

where $\chi^2_r(1 - \alpha)$ denotes the $1 - \alpha$ quantile of the χ^2-distribution with r degrees of freedom.

With probability $1 - \alpha$, $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i^{T_i}$ belongs to $\mathcal{E}_\alpha(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$. (Jennrich, 1969, Theorem 6) is used to get the strong consistency and (Jennrich, 1969, Theorem 7) is used to get the asymptotic normality of $\hat{\boldsymbol{\theta}}_i^{T_i}$. □

B.2 Proof of Theorem 4

Proof. We use the dominated convergence theorem. As the variance is lower bounded by Σ^θ, the mean is upper bounded by a fixed constant, so we can construct a dominating function. Then,

$$\begin{aligned}
\lim_{\sigma_\varepsilon^2 \to 0} m_{\hat{\boldsymbol{\theta}}_i^\infty}(\eta) &= \lim_{\sigma_\varepsilon^2 \to 0} \int_{\mathbb{R}^r} \varphi(\boldsymbol{\theta}_i; c, C) \varphi(\eta; \boldsymbol{\theta}_0, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) + \Sigma^{\boldsymbol{\theta}}) d\boldsymbol{\theta}_i \\
&= \int_{\mathbb{R}^r} \lim_{\sigma_\varepsilon^2 \to 0} \left(\varphi(\boldsymbol{\theta}_i; c, C) \varphi(\eta; \boldsymbol{\theta}_0, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) + \Sigma^{\boldsymbol{\theta}})\right) d\boldsymbol{\theta}_i \\
&= \int_{\mathbb{R}^r} \varphi(\boldsymbol{\theta}_i; \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}) \varphi(\eta; \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}) d\boldsymbol{\theta}_i = \varphi(\eta; \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}).
\end{aligned}$$

□

B.3 Proof of Theorem 5

We consider $\eta \in \mathcal{E}_\rho(\boldsymbol{\theta}_0, \mathbf{A})$, with $\mathbf{A}$ positive definite.

$$\begin{aligned}
m_{\hat{\boldsymbol{\theta}}_i^\infty}(\eta) &= \int_{\mathbb{R}^r} \varphi\left(\boldsymbol{\theta}_i; \eta, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_i)\right) \varphi\left(\boldsymbol{\theta}_i; \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}\right) d\boldsymbol{\theta}_i \\
&= \int_{\mathbb{R}^r} \varphi\left(\boldsymbol{\theta}_i; \eta, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0)\right) \varphi\left(\boldsymbol{\theta}_i; \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}\right) \times \det(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_0) \mathbf{a}_i(\boldsymbol{\theta}_i)) \\
&\qquad \times \exp\left(-\frac{1}{2\sigma_\varepsilon^2} (\boldsymbol{\theta}_i - \eta)^T \left(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) - \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0)\right) (\boldsymbol{\theta}_i - \eta)\right) d\boldsymbol{\theta}_i \\
&= \varphi(\eta; \boldsymbol{\theta}_0, \sigma_\varepsilon^2 \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0) + \Sigma^{\boldsymbol{\theta}}) \int_{\mathbb{R}^r} \varphi(\boldsymbol{\theta}_i; c_0, C_0) \times \det(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_0) \mathbf{a}_i(\boldsymbol{\theta}_i)) \\
&\qquad \times \exp\left(-\frac{1}{2\sigma_\varepsilon^2} (\boldsymbol{\theta}_i - \eta)^T \left(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) - \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0)\right) (\boldsymbol{\theta}_i - \eta)\right) d\boldsymbol{\theta}_i
\end{aligned}$$

where the last equality comes from Lemma 4, with

$$\begin{aligned}
C_0 &= (\sigma_\varepsilon^{-2} \mathbf{a}_i(\boldsymbol{\theta}_0) + (\Sigma^{\boldsymbol{\theta}})^{-1})^{-1}; \\
c_0 &= C_0 \left(\sigma_\varepsilon^{-2} \mathbf{a}_i(\boldsymbol{\theta}_0) \eta + (\Sigma^{\boldsymbol{\theta}})^{-1} \boldsymbol{\theta}_0\right) = \boldsymbol{\theta}_0 + C_0 \sigma_\varepsilon^{-2} \mathbf{a}_i(\boldsymbol{\theta}_0)(\eta - \boldsymbol{\theta}_0).
\end{aligned}$$

Our goal is now to prove that

$$\mathcal{I} = \int_{\mathbb{R}^r} \varphi(\boldsymbol{\theta}_i; c_0, C_0) \det(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_0)\mathbf{a}_i(\boldsymbol{\theta}_i)) \times \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{\theta}_i - \eta)^T \left(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) - \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0)\right)(\boldsymbol{\theta}_i - \eta)\right) d\boldsymbol{\theta}_i$$

is close to 1.

We then divide this integral into two parts: let $\tilde{\rho} > 0$, if $\boldsymbol{\theta}_i \in \mathcal{E}_{\tilde{\rho}}(\boldsymbol{\theta}_0, A)$,

$$\begin{aligned} \mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) &= \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0) + D_{\boldsymbol{\theta}_0}\mathbf{a}_i^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0) + O(\tilde{\rho}), \\ \Rightarrow (\boldsymbol{\theta}_i - \eta)^T(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) - \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0))(\boldsymbol{\theta}_i - \eta) &= O(\tilde{\rho} + \rho), \text{ and} \\ \det(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_0)\mathbf{a}_i(\boldsymbol{\theta}_i)) &= 1 + O(\tilde{\rho}). \end{aligned}$$

If $\boldsymbol{\theta}_i \in \mathcal{E}_{\tilde{\rho}}(\boldsymbol{\theta}_0, A)^c$, by Assumption E we get that

$$\begin{aligned} \det(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_0)\mathbf{a}_i(\boldsymbol{\theta}_i)) &= O(1) \\ \Rightarrow \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{\theta}_i - \eta)^T \left(\mathbf{a}_i^{-1}(\boldsymbol{\theta}_i) - \mathbf{a}_i^{-1}(\boldsymbol{\theta}_0)\right)(\boldsymbol{\theta}_i - \eta)\right) &\leq 1. \end{aligned}$$

This leads to

$$\begin{aligned} \mathcal{I} &= (1 + O(\rho)) \int_{\mathcal{E}_{\tilde{\rho}}(\boldsymbol{\theta}_0, A)} \varphi(\boldsymbol{\theta}_i; c_0, C_0) d\boldsymbol{\theta}_i + O(1) \int_{\mathcal{E}_{\tilde{\rho}}(\boldsymbol{\theta}_0, A)^c} \varphi(\boldsymbol{\theta}_i; c_0, C_0) d\boldsymbol{\theta}_i \\ &= 1 + O(\rho). \end{aligned}$$

□

B.4 Proof of Theorem 6

From Theorem 4,

$$\sqrt{T_i}(\hat{\boldsymbol{\theta}}_i^{T_i} - \boldsymbol{\theta}_0) \underset{T_i \to +\infty}{\overset{d}{\to}} \mathcal{N}_r(\mathbf{0}, \Sigma^{\boldsymbol{\theta}}).$$

For $i = 1, \ldots, n$, let $Z_i^{T_i} \sim \mathcal{N}(\boldsymbol{\theta}_0, \frac{1}{T_i}\Sigma^{\boldsymbol{\theta}})$, and $\Delta_i^{T_i} = \hat{\boldsymbol{\theta}}_i^{T_i} - Z_i^{T_i}$.

We know that $\Delta_i^{T_i} \underset{T_i \to +\infty}{\overset{d}{\to}} \delta_0$ with δ_0 the Dirac distribution in 0, which implies that $\Delta_i^{T_i} \underset{T_i \to +\infty}{\overset{P}{\to}} 0$. Then, by inverting limits, $\frac{1}{n}\sum_{i=1}^n \Delta_i^{T_i} \underset{T_i \to +\infty, n \to +\infty}{\overset{d}{\to}} \delta_0$.

For $Z_i^{T_i}$, we use Lindeberg Central limit Theorem, recalled in Theorem A of Appendix C.
Let $S_n^2 = \Sigma^{\boldsymbol{\theta}} \sum_{i=1}^n \frac{1}{T_i}$. We assume that (14) holds. Then,

$$b_{n,\mathbf{T}}^{-1/2}(\frac{1}{n}\sum_{i=1}^n Z_i^{T_i} - \boldsymbol{\theta}_0) \underset{n \to +\infty}{\overset{d}{\to}} \mathcal{N}(0, \Sigma^{\boldsymbol{\theta}}).$$

By Slustky's Lemma, we get that

$$b_{n,\mathbf{T}}^{-1/2}\left(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0\right) \underset{n \to +\infty, \min T_i \to +\infty}{\overset{d}{\to}} \mathcal{N}_r(\mathbf{0}, \Sigma^{\boldsymbol{\theta}}).$$

For $i = 1, \ldots, n$, let $Z_i^{T_i} \sim \mathcal{N}(\boldsymbol{\theta}_0, \frac{1}{T_i}\Sigma^{\boldsymbol{\theta}})$, and $\tilde{\Delta}_i^{T_i} = \sqrt{T_i}(\hat{\boldsymbol{\theta}}_i^{T_i}\hat{\boldsymbol{\theta}}_i^{T_i} - Z_i^{T_i})$. We know that $\sqrt{T_i}Z_i^{T_i} \sim \mathcal{N}(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ and then $\sum_{i=1}^n T_i Z_i^{T_i}(Z_i^{T_i})^T \sim \mathcal{W}(\Sigma^{\boldsymbol{\theta}}, n-1)$. Moreover, we know that $\tilde{\Delta}_i^{T_i} \underset{T_i \to +\infty}{\overset{d}{\to}} \delta_0$,and then

$$\frac{1}{n-1}\sum_{i=1}^n T_i \Delta_i^{T_i}(\Delta_i^{T_i})^T \underset{\min T_i \to +\infty, n \to +\infty}{\overset{d}{\to}} \delta_0.$$

Thus, we get

$$\begin{aligned} \frac{1}{n-1}\sum_{i=1}^n T_i \hat{\boldsymbol{\theta}}_i^{T_i}(\hat{\boldsymbol{\theta}}_i^{T_i})^T = &\frac{1}{n-1}\sum_{i=1}^n T_i Z_i^{T_i}(Z_i^{T_i})^T + \frac{1}{n}\sum_{i=1}^n T_i \Delta_i^{T_i}(\Delta_i^{T_i})^T \\ &\underset{n \to +\infty, \min T_i \to +\infty}{\overset{d}{\to}} \mathcal{W}(\Sigma^{\boldsymbol{\theta}}, n-1). \end{aligned}$$

□

B.5 Proof of Theorem 7

We check that our model satisfies the assumptions given in (Pinheiro, 1994, Chapter 3):

- The U_i are independent and follow a $\mathcal{N}(0, \Sigma^{U_i})$ distribution, ε_i follows a $\mathcal{N}_{T_i}(0, \sigma_\varepsilon^2)$ distribution and the U_i are independent of ε_i
- The matrix $\mathbf{B}_i^\mu$ is of full rank, as it is a functional basis
- $n \geq m_\mu + 1 + 1$
- The concatenated matrix $[\mathbf{B}_i^\mu, \mathbf{B}_i^U]$ has rank greater than m_μ if we don't take the same basis for μ and U_i
- The matrices I_{T_i} and $\mathbf{B}_i^U(\mathbf{B}_i^U)^t$ are linearly independent
- $\lim_{n\to+\infty} \frac{n - rank(\mathbf{B}_i^U)}{n} = 1$

So we get the asymptotic normality of the estimator. □

B.6 Proof of Theorem 8

Recall the different operator parts of the iterative algorithm in (12). The Banach fixed point theorem, recalled in Theorem B of Appendix C, is used to prove that there is a unique fixed point, and that the algorithm converges. To use this theorem, we work in $\mathbb{R}^{m_\mu + nm_U + 1 + nr}$ with the Euclidean distance. It is a non-empty complete metric space. The mapping we consider is Ψ, as defined in (12). We need to prove that Ψ is a contraction mapping.

Denote by k_f the Lipschitz constant for the function f.

We want to find k_Ψ such that, for $(x, y) \in (\mathbb{R}^{m_\mu + nm_U + 1 + nr})^2$,

$$\|\Psi(x) - \Psi(y)\|_2 \leq k_\Psi \|x - y\|_2.$$

As Ψ_2 and Ψ_3 are linear, for $(x, y) \in (\mathbb{R}^{m_\mu + nm_U + 1 + nr})^2$,

$$\|\Psi_3 \circ \Psi_2 \circ \Psi_1(x) - \Psi_3 \circ \Psi_2 \circ \Psi_1(y)\|_2 = \|\Psi_3 \circ \Psi_2(\Psi_1(x) - \Psi_1(y))\|_2$$

The proof relies on Lemma 5 applied to Ψ_3 and Ψ_2, defined via the mixed effect models. The statement of Lemma 5 and its proof can be found in Appendix C. Then, using Assumption G, there exists $k_\Psi < 1$ such that Ψ is k_Ψ-Lipschitz. Banach fixed point theorem concludes. □

B.7 Proof of Theorem 9

First, we prove that for a fixed iteration ite, the several computations we are doing to the true parameters $(\boldsymbol{\alpha}^\mu, \sigma_{\boldsymbol{\varepsilon}}^2, \underset{\sim}{\Sigma}^U, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ are keeping it fixed under Assumption C.

Fix the iteration number (ite) and consider $(\boldsymbol{\alpha}^\mu, \sigma_{\boldsymbol{\varepsilon}}^2, \underset{\sim}{\Sigma}^U, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$. We know $(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ and we predict $\boldsymbol{\theta}_i$ with the BLUP. As $n \to \infty$ and $\min_i T_i \to \infty$, predictions are good. Then, we estimate $(\boldsymbol{\alpha}^\mu, \sigma_{\boldsymbol{\varepsilon}}^2, \underset{\sim}{\Sigma}^U)$, by Theorem 7 we get $(\boldsymbol{\alpha}^\mu, \sigma_{\boldsymbol{\varepsilon}}^2, \underset{\sim}{\Sigma}^U)$ (strong consistency).

Then, we approximate $\boldsymbol{\theta}_i$ for all $i = 1, \ldots, n$. By Theorem 4, those pseudo-observations are close to the true random variables, with the good distribution function. Finally, we estimate a linear mixed model on those observations: by Theorem 6, we get $\boldsymbol{\theta}_0$ and $\Sigma^{\boldsymbol{\theta}}$ (consistency).

Then, under identifiability, Assumptions A, C and G, $(\boldsymbol{\alpha}^\mu, \sigma_{\boldsymbol{\varepsilon}}^2, \underset{\sim}{\Sigma}^U, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ is a fixed point of Ψ.

By Theorem 8, there exists only one fixed point: then

$$((\hat{\boldsymbol{\alpha}}^\mu)^{(\infty)}, (\hat{\sigma}_\varepsilon)^{(\infty)}, (\hat{\underset{\sim}{\Sigma}}^U)^{(\infty)}, \hat{\boldsymbol{\theta}}_0^{(\infty)}, (\hat{\Sigma}^{\boldsymbol{\theta}})^{(\infty)}) \underset{\substack{n\to\infty \\ \min T_i \to \infty}}{\longrightarrow} (\boldsymbol{\alpha}^\mu, \sigma_{\boldsymbol{\varepsilon}}^2, \underset{\sim}{\Sigma}^U, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}).$$

This convergence is almost surely, as the convergence in each step is almost surely. □

C Additional useful results and tools.

Theorem A (Lindeberg Central Limit Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $X_k : \Omega \to \mathbb{R}$, $k \in \mathbb{N}$, be independent random variables defined on that space. Assume that the expected values $E[X_k] = \mu_k$ and variances $\mathrm{Var}[X_k] = \sigma_k^2$ exist and are finite. Define $s_n^2 = \sum_{k=1}^n \sigma_k^2$.*

If this sequence of independent random variables X_k satisfies Lindeberg's condition: for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} E\left[(X_k - \mu_k)^2 \cdot \mathbf{1}_{\{|X_k - \mu_k| > \varepsilon s_n\}}\right] = 0, \tag{14}$$

where $\mathbf{1}$ is the indicator function, then the central limit theorem holds, i.e. the random variables $Z_n := \frac{\sum_{k=1}^n (X_k - \mu_k)}{s_n}$ converge in distribution to a standard normal random variable as $n \to \infty$.

Theorem B (Banach fixed point theorem)**.** *Let (X, d) be a non-empty complete metric space with a contraction mapping $T : X \to X$ with Lipschitz constant $q \in [0, 1)$. Then, T admits a unique fixed-point x^∞ in X. Furthermore, x^∞ can be found as follows: starts with an arbitrary element x_0 in X and define a sequence $\{x_n\}$ by $x_n = T(x_{n-1})$, then $x_n \to x^\infty$. Moreover, $d(x^\infty, x_n) = q^n d(x_1, x_0)/(1-q)$.*

Lemma 4 (Ahrendt (2005))**.** *The following holds, for all $x \in \mathbb{R}^p$, for A, B positive definite matrices of size $p \times p$ and $(a, b) \in (\mathbb{R}^p)^2$:*

$$\varphi(x; a, A)\varphi(x; b, B) = \varphi(a; b, A + B)\varphi(x; c, C),$$

with $C = (A^{-1} + B^{-1})^{-1}$ and $c = C(A^{-1}a + B^{-1}b)$.

Lemma 5. *Let π_1 and π_2 be two orthogonal projections, and denote $|||.|||$ the operator norm. Then,*

$$|||\pi_1||| = |||\pi_2||| = 1; \tag{15}$$

$$|||\pi_1 \circ \pi_2||| < 1 \text{ if and only if } E_{\pi_1} \cap E_{\pi_2} = \{0\}. \tag{16}$$

Proof. We first prove (15). Let π be an orthogonal projection. As it is a projection, its norm is larger than 1. As it is an orthogonal projection, we can use Pythagorean theorem to prove that its norm is smaller than 1.

We now prove (16). If $E_{\pi_1} \cap E_{\pi_2} \neq \{0\}$, let $x \in E_{\pi_1} \cap E_{\pi_2}$. Then, $\pi_1 \circ \pi_2(x) = x$ so $|||\pi_1 \circ \pi_2||| \geq 1$.

If $E_{\pi_1} \cap E_{\pi_2} = \{0\}$, assume that $|||\pi_1 \circ \pi_2||| = 1$: there exists $x \neq 0$ such that $\|\pi_1 \circ \pi_2(x)\| = \|x\|$. But as π_1 and π_2 are projections, it means that $x \in E_{\pi_1} \cap E_{\pi_2}$: contradiction. □

Acknowledgements

The authors thank Professor Alois Kneip for helpful discussions, and Professor Daniel Gervini for help with his computer codes. E. Devijver also sincerely thanks R. Molinier for fruitful discussions.

References

Ahrendt, P. (2005). The multivariate gaussian probability distribution. Technical report. Available at http://orbit.dtu.dk/files/2691604/imm3312.pdf.

Bigot, J. (2013). Fréchet means of curves for signal averaging and application to ECG data analysis. *Ann. Appl. Stat.*, 7(4):2384–2401.

Carroll, C., Müller, H.-G., and Kneip, A. (2020). Cross-component registration for multivariate functional data, with application to growth curves. *Biometrics*, to appear.

Chakraborty, A. and Panaretos, V. M. (2021). Functional registration and local variations: identifiability, rank, and tuning. *Bernoulli*, 27(2):1103–1130.

Chen, H. and Wang, Y. (2011). A penalized spline approach to functional mixed effects model analysis. *Biometrics*, 67(3):861–870.

Cheng, W., Dryden, I. L., and Huang, X. (2016). Bayesian registration of functions and curves. *Bayesian Anal.*, 11(2):447–475.

Claeskens, G., Silverman, B. W., and Slaets, L. (2010). A multiresolution approach to time warping achieved by a Bayesian prior-posterior transfer fitting strategy. *Journal of the Royal Statistical Society Series B*, 72(5):673–694.

Dupuy, J.-F., Loubes, J.-M., and Maza, E. (2011). Non parametric estimation of the structural expectation of a stochastic increasing function. *Statistics and Computing*, 21(1):121–136.

Elmi, A., Ratcliffe, S., Parry, S., and Guo, W. (2011). A B-spline based semiparametric nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, 20(2):492–509.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice*. Springer Series in Statistics. Springer.

Gervini, D. (2015a). Dynamic retrospective regression for functional data. *Technometrics*, 57(1):26–34.

Gervini, D. (2015b). Warped functional regression. *Biometrika*, 102(1):1–14.

Gervini, D. and Carter, P. A. (2014). Warped functional analysis of variance. *Biometrics*, 70(3):526–535.

Gervini, D. and Gasser, T. (2004). Self-modelling warping functions. *Journal of the Royal Statistical Society: Series B*, 66(4):959–971.

Guo, W. (2002). Functional mixed effects models. *Biometrics*, 58(1):121–128.

Hadjipantelis, P. Z., Aston, J. A. D., Müller, H.-G., and Moriarty, J. (2014). Analysis of spike train data: A multivariate mixed effects model for phase and amplitude. *Electron. J. Statist.*, 8(2):1797–1807.

Happ, C., Scheipl, F., Gabriel, A.-A., and Greven, S. (2019). A general framework for multivariate functional principal component analysis of amplitude and phase variation. *Stat*, 8(1).

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40(2):633–643.

Kneip, A. and Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.*, 20(3):1266–1305.

Marron, J. S., Ramsay, J. O., Sangalli, L. M., and Srivastava, A. (2015). Functional data analysis of amplitude and phase variation. *Statist. Sci.*, 30(4):468–484.

Park, J. and Ahn, J. (2017). Clustering multivariate functional data with phase variation. *Biometrics*, 73(1):324–333.

Pinheiro, J. C. (1994). Topics in mixed effects models. http://www.math.ku.dk/ erhansen/web/stat1/pinheiro.pdf.

Rakêt, L. L., Sommer, S., and Markussen, B. (2014). A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data. *Pattern Recognition Letters*, 38:1 – 7.

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer-Verlag.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer-Verlag, second edition.

Ramsay, J., Wang, X., and Flanagan, R. (1995). A functional data analysis of the pinch force of human fingers. *Appl. Statist.*, 44(1):17–30.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010). k-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219 – 1233.

Strait, J., Kurtek, S., Bartha, E., and MacEachern, S. N. (2017). Landmark-constrained elastic shape analysis of planar curves. *Journal of the American Statistical Association*, 112(518):521–533.

Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334.

Tang, L., Zeng, P., Shi, J. Q., and Kim, W.-S. (2020). Joint curve registration and classification with two-level functional models. *arXiv*, 2011.02304.

Tucker, J. D., Wu, W., and Srivastava, A. (2013). Generative models for functional data using phase and amplitude separation. *Comput. Stat. Data Anal.*, 61(C):50–66.

Wagner, H. and Kneip, A. (2019). Nonparametric registration to low-dimensional function spaces. *Computational Statistics & Data Analysis*, 138:49 – 63.

Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. *Ann. Statist.*, 25:1251–1276.

Wang, W. (2013). Identifiability of linear mixed effects models. *Electron. J. Statist.*, 7:244–263.

Wrobel, J., Zipunnikov, V., Schrack, J., and Goldsmith, J. (2019). Registration for exponential family functional data. *Biometrics*, 75(1):48–57.

Xie, W., Kurtek, S., Bharath, K., and Sun, Y. (2017). A geometric approach to visualization of variability in functional data. *Journal of the American Statistical Association*, 112(519):979–993.

Yu, Q., Lu, X., and Marron, J. S. (2017). Principal nested spheres for time-warped functional data analysis. *Journal of Computational and Graphical Statistics*, 26(1):144–151.

Zeng, P., Shi, J. Q., and Kim, W.-S. (2019). Simultaneous registration and clustering for multidimensional functional data. *Journal of Computational and Graphical Statistics*, 28(4):943–953.