HAL open science

A robust algorithm for template curve estimation based on manifold embedding

Chloé Dimeglio, Santiago Gallón, Jean-Michel Loubes, Elie Maza

To cite this version:

Chloé Dimeglio, Santiago Gallón, Jean-Michel Loubes, Elie Maza. A robust algorithm for template curve estimation based on manifold embedding. 2013. hal-00834080v3

HAL Id: hal-00834080

https://hal.science/hal-00834080v3

Submitted on 14 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A robust algorithm for template curve estimation based on manifold embedding

Chloé Dimeglio[*a], Santiago Gallón[†a], Jean-Michel Loubes[‡a], and Elie Maza[§c]

[a] *Institut de Mathématiques de Toulouse, Université Toulouse III Paul Sabatier, Toulouse, France.*

[c] *Laboratoire Génomique et Biotechnologie des Fruits, UMR 990 INRA/INP-ENSAT, Université Toulouse III Paul Sabatier, Toulouse, France.*

June 14, 2013

Abstract

This paper considers the problem of finding a meaningful template function that represents the common pattern of a sample of curves. To address this issue, a novel algorithm based on a robust version of the isometric featuring mapping (Isomap) algorithm is developed. Assuming that the functional data lie on an intrinsically low-dimensional smooth manifold with unknown underlying structure, we propose an approximation of the geodesic distance. This approximation is used to compute the corresponding empirical Fréchet median function, which provides an intrinsic estimator of the template function. Unlike the Isomap method, the algorithm has the advantage of being parameter free and easier to use. Comparisons with other methods, with both simulated and real datasets, show that the algorithm works well and outperforms these methods.

Key words: Fréchet median; functional data analysis; Isomap.

1 Introduction

Nowadays, experiments where the outcome constitutes a sample of functions $\{f_i(t) \colon t \in \mathcal{T} \subset \mathbb{R},\, i = 1, \ldots, n\}$ are more and more frequent. Such kind of functional data are now commonly encountered in speech signal recognition in engineering, growth curves analysis in biology and medicine, microarray experiments in molecular biology and genetics, expenditure and income studies in economics, just to name a few.

However, extracting the information conveyed by all the curves is a difficult task. Indeed when finding a meaningful representative function that characterizes the common behavior of the sample, capturing its inner characteristics (as trends, local extrema and inflection points), a major difficulty comes from the fact that

[*] cd@geosys.com

[†] santiagog@udea.edu.co

[‡] jean-michel.loubes@math.univ-toulouse.fr

[§] Elie.Maza@ensat.fr

usually there are both amplitude (variation on the y-axis) and phase (variation on the x-axis) variations with respect to the common pattern, as pointed out in Ramsay and Li [27], Ramsay and Silverman [28], or Vantini [37] for instance. Hence, in the two last decades, there has been a growing interest for statistical methodologies and algorithms to remove the phase variability and recover a single template conveying all the information in the data since the classical cross-sectional mean is not a good representative of the data (see for instance Kneip and Gasser [19]).

Two different kinds of methods have been developed for template function estimation. The first group relies on the assumption that there exists a mean pattern from which all the observations differ, i.e an unknown function f such that each observed curve is given by $f_i(t) = f \circ h_i(t)$, where h_i are deformation functions. Hence finding this patten is achieved by aligning all curves f_i. This method is known as *curve registration*. In this direction, various curve registration methods have been proposed using different strategies. When the warping operator is not specified, we refer for instance to Kneip and Gasser [19], Wang and Gasser [38] Kneip et al. [21], James [17], Tang and Müller [34], and Kneip and Ramsay [20] or Dupuy et al. [14]. When a parametric model for the deformation is chosen, the statistical problem requires a semi-parametric approach through a self-modeling regression framework $f_i(t) = f(t, \theta_i)$ (see Kneip and Gasser [18]), where all functions are deduced with respect to the template f by mean a finite-dimensional individual parameter vector θ_i. This point of view is also followed in Silverman [33], Rønn [29], Gamboa et al. [16], Castillo and Loubes [6], Bigot et al. [5] and Trigano et al. [36].

The second category of methods do not assume any deformation model for the individual functions. The purpose is to select a curve which is assumed to be located at the *center* of the functions and estimate it directly from the data without stressing any particular curve. This is achieved for instance by López-Pintado and Romo [24] and Arribas-Gil and Romo [1] estimating the template based on the concept of depth for functional data as measure of centrality of the sample.

In this paper, we propose an alternative way based on the ideas of manifold learning theory. We assume that the observed functions can be modeled as variables with values in a manifold $\mathcal{M}$ with an unknown geometry. Although the manifold is unknown, the key property is that its underlying geometric structure is contained in the sample of observed curves so that the geodesic distance can be reconstructed directly from the data. The template curve estimation is then equivalent to consider a location measure of the data with respect to this geodesic distance, hence approximating the Fréchet mean or median of the data. Recently, Chen and Müller [7] have also adopted a similar methodology appealing to the nonlinear manifold representation for functional data. Several algorithms have been developed over the last decade in order to reconstruct the natural embedding of data onto a manifold. Some of these are, for instance, the Isometric featuring mapping −Isomap− (Tenenbaum et al. [35]), Local Linear Embedding −LLE− (Roweis and Saul [30]), Laplacian Eigenmap (Belkin and Niyogi [3]), Hessian Eigenmap (Donoho and Grimes [13]), Diffusion maps (Coifman and Lafon [8]), Local Tangent Space Alignment −LTSA− (Zhang and Zha [39]), among others. In the following, we propose a robust version of the Isomap algorithm devoted to

functional data, less sensitive to outliers and easier to handle. The performance of the algorithm is evaluated both on simulations and real data sets.

This article is organized as follows. The frame of our study is described in Section 2. Section 3 is devoted to the robust modification of the Isomap algorithm proposed to the metric construction of the approximated geodesic distance based on the observed curves. In Section 4 we analyze the template estimation problem in a shape invariant model, showing that this issue can be solved using the manifold geodesic approximation procedure. In Section 5, the performance of our algorithm is studied using simulated data. In Section 6, several applications on real functional data sets are performed. Some concluding remarks are given in Section 7.

2 Template estimation with a manifold embedding framework

Consider discrete realizations of functions f_i observed at time $t_j \in \mathcal{T}$, with $\mathcal{T}$ a bounded interval of $\mathbb{R}$. Set $X_i = \{f_i(t_j),\ j = 1, \ldots, m\} \in \mathbb{R}^m$ for $i = 1, \ldots, n$. We assume that the data have a common structure which can be modeled as a manifold embedding. Hence the sample $\mathcal{E} = \{X_1, \ldots, X_n\}$ consists of i.i.d random variables sampled from a law $Q \in \mathcal{M}$,where $\mathcal{M}$ is an unknown connected smooth submanifold of $\mathbb{R}^m$, endowed with the geodesic distance δ induced by the Riemannian metric g on $\mathcal{M} \subset \mathbb{R}^m$ (see for instance do Carmo [12]).

Under this geometrical framework, the statistical analysis of the curves should be carried out carefully, using the intrinsic geodesic distance and not the Euclidean distance, see for instance Pennec [26] . In particular, an extension of the usual notion of central value from Euclidean spaces to arbitrary manifolds is based on the Fréchet function, defined by

Definition 1 (Fréchet function)**.** Let $(\mathcal{M}, \delta)$ be a metric space and let $\alpha > 0$ be a given real number. For a given probability measure Q defined on the Borel σ-field of $\mathcal{M}$, the Fréchet function of Q is given by

$$F_\alpha(\mu) = \int_{\mathcal{M}} \delta^\alpha(X, \mu) Q(\mathrm{d}x), \qquad \mu \in \mathcal{M}.$$

For $\alpha = 1$ and $\alpha = 2$, the minimizers of $F_\alpha(\mu)$, if there exist, are called the Fréchet (or intrinsic) median and mean respectively. Following Koenker [22], in this paper we will particularly deal with the intrinsic median, denoted by $\mu_I^1(Q)$ to obtain a robust estimate for the template function $f \in \mathcal{M}$. Hence define the corresponding empirical intrinsic median as

$$\widehat{\mu}_I^1 = \arg\min_{\mu \in \mathcal{M}} \sum_{i=1}^{n} \delta\left(X_i, \mu\right). \tag{1}$$

However, the previous estimator relies on the unobserved manifold $\mathcal{M}$ and its underlying geodesic distance δ. A popular estimator is given by the Isomap algorithm

for δ. The idea is to build a simple metric graph from the data, which will be close enough from the manifold. Hence the approximation of the geodesic distance between two points depends on the length of the edges of the graph which connect these points. The algorithm approximates the unknown geodesic distance δ between all pairs of points in $\mathcal{M}$ in terms of shortest path distance between all pairs of points in a nearest neighbor graph $\mathcal{G}$ constructed from the data points $\mathcal{E}$. If the discretization of the manifold contains enough points with regards to the curvature of the manifold, hence the graph distance will be a good approximation of the geodesic distance. For details about the Isomap algorithm, see Tenenbaum et al. [35], Bernstein et al. [4], and de Silva and Tenenbaum [11].

The construction of the weighted neighborhood graph in the first step of the Isomap algorithm requires the choice of a parameter which controls the neighborhood size and therefore its success. This is made according to a $K-$rule (connecting each point with its K nearest neighbors) or $\epsilon-$rule (connecting each point with all points lying within a ball of radius ϵ) which are closely related to the local curvature of the manifold. Points which are too distant to be connected to the biggest graph are not used, making the algorithm unstable (see Balasubramanian and Schwartz [2]). In this paper we propose a robust version of this algorithm which leads to an approximation of the geodesic distance, $\hat{\delta}$. Our version does not exclude any point and does not require any additional tuning parameter. This algorithm has been applied with success to align density curves in microarray data analysis (task known as normalization in bioinformatics) in Gallón et al. [15]. The construction of the approximated geodesic distance is detailed in Section 3.

Once an estimator of the geodesic distance is built, we propose to estimate the empirical Fréchet median by its approximated version

$$\widehat{\mu}_{I,n}^{1} = \arg\min_{\mu \in \mathcal{G}} \sum_{i=1}^{n} \hat{\delta}\left(X_i, \mu\right). \tag{2}$$

This estimator is restricted to stay within the graph $\mathcal{G}$ since the approximated geodesic distance is only defined on the graph. Hence we choose as a pattern of the observation the point which is at the *center* of the dataset, where center has to be understood with respect to the inner geometry of the observations.

3 The robust manifold learning algorithm

Let X be a random variable with values in an unknown connected and geodesically complete Riemannian manifold $\mathcal{M} \subset \mathbb{R}^m$, and a sample $\mathcal{E} = \{X_i \in \mathcal{M},\ i = 1, \ldots, n\}$ with distribution Q. Set d the Euclidean distance on $\mathbb{R}^m$ and δ the induced geodesic distance on $\mathcal{M}$. Our aim is to estimate the geodesic distance between two points on the manifold $\delta\left(X_i, X_{i'}\right)$ for all $i \neq i' \in \{1, \ldots, n\}$.

The Isomap algorithm proposes to learn the manifold topology from a neighborhood graph. In the same way, our purpose is to approximate the geodesic distance δ between

a pair of data points by the graph distance on the shortest path between the pair on the neighborhood graph. The main difference between our algorithm and the Isomap algorithm lies in the treatment of points which are far from the others. Indeed, the first step of the original Isomap algorithm consists in constructing the K-nearest neighbor graph or the ϵ-nearest neighbor graph for a given positive integer K or a real $\epsilon > 0$, respectively and then to exclude points which are not connected to the graph. Such a step is not present in our algorithm since we consider that a distant point is not always considered an outlier. Hence, we do not exclude any points. Moreover, a sensitive issue of the Isomap algorithm is that it requires the choice of the neighbor parameter (K or ϵ) which is closely related to the local curvature of the manifold, determining the quality of the construction (see, for instance, Balasubramanian and Schwartz [2]). In our algorithm, we give a tuning parameter free process to simplify the analysis.

The algorithm has three steps. The first step constructs a complete weighted graph associated to $\mathcal{E}$ based on Euclidean distances $\mathrm{d}(X_i, X_{i'})$ between all pairwise points $X_i, X_{i'} \in \mathbb{R}^m$. It is a complete Euclidean graph $\mathcal{G}_{\mathrm{E}} = (\mathcal{E}, E)$ with set of nodes $\mathcal{E}$ and set of edges $E = \{\{X_i, X_{i'}\},\ i = 1, \ldots, n-1,\ i' = i+1, \ldots, n\}$ weighted with the corresponding Euclidean distances.

In the second step, the algorithm obtains the Euclidean Minimum Spanning Tree $\mathcal{G}_{\mathrm{MST}} = (\mathcal{E}, E_{\mathrm{T}})$ associated to $\mathcal{G}_{\mathrm{E}}$, i.e. the spanning tree that minimizes the sum of the weights of the edges in the spanning tree of $\mathcal{G}_{\mathrm{E}}$, $\sum_{\{X_i, X_{i'}\} \in E_{\mathrm{T}}} \mathrm{d}\left(X_i, X_{i'}\right)$. The underlying idea in this construction is that, if two points X_i and $X_{i'}$ are relatively close, then we have that $\delta\left(X_i, X_{i'}\right) \approx \mathrm{d}\left(X_i, X_{i'}\right)$. This may not be true if the manifold is very twisted and/or if too few points are observed, and may induce bad approximations. So the algorithm will produce a good approximation for relatively regular manifolds. This drawback is well known when dealing with graph-based approximations of the geodesic distance (Tenenbaum et al. [35], and de Silva and Tenenbaum [11]).

An approximation of $\delta\left(X_i, X_{i'}\right)$ is provided by the sum of all the Euclidean distances of the edges of the shortest path on $\mathcal{G}_{\mathrm{MST}}$ connecting X_i to $X_{i'}$, i.e. $\hat{\delta}\left(X_i, X_{i'}\right) = \min_{g_{ii'} \in \mathcal{G}_{\mathrm{MST}}} L\left(g_{ii'}\right)$, where $L\left(g_{ii'}\right)$ denotes the length of a path $g_{ii'}$ connecting X_i to $X_{i'}$ on $\mathcal{G}_{\mathrm{MST}}$. However, this construction is highly unstable since the addition of new points may change completely the structure of the graph.

To cope with this problem, we propose in the third stage to add more robustness in the construction of the approximation graph. Actually, in our algorithm we add more edges between the data points to add extra paths and thus to cover better the manifold. The underlying idea is that paths which are close to the ones selected in the construction of the $\mathcal{G}_{\mathrm{MST}}$ could also provide good alternate ways of connecting the edges. Closeness here is understood as lying in open balls $B\left(X_i, \epsilon_i\right) \subset \mathbb{R}^m$ around the point X_i with radius $\epsilon_i = \max_{\{X_i, X_i\} \in E_{\mathrm{T}}} \mathrm{d}\left(X_i, X_{i'}\right)$. Hence, these new paths between the data are admissible and should be added to the edges of the graph. Finally, we obtain a new robustified graph $\mathcal{G}' = (\mathcal{E}, E')$ defined by

$$\{X_i, X_{i'}\} \in E' \Longleftrightarrow \overline{X_i X_{i'}} \subset \bigcup_{i=1}^{n} B\left(X_i, \epsilon_i\right),$$

where

$$\overline{X_i X_{i'}} = \{X \in \mathbb{R}^m,\ \exists \lambda \in [0,1],\ X = \lambda X_i + (1-\lambda) X_{i'}\}\,.$$

Finally, $\mathcal{G}'$ is the graph which gives rise to our estimator of δ, given by

$$\hat{\delta}\left(X_i, X_{i'}\right) = \min_{g_{ii'} \in \mathcal{G}'} L\left(g_{ii'}\right). \tag{3}$$

Hence, $\hat{\delta}$ is the distance associated with $\mathcal{G}'$, that is, for each pair of points X_i and $X_{i'}$, we have $\hat{\delta}\left(X_i, X_{i'}\right) = \mathrm{L}\left(\hat{\gamma}_{ii'}\right)$ where $\hat{\gamma}_i$ is the minimum length path between X_i and $X_{i'}$ associated to $\mathcal{G}'$. We point out that all points of the data sets are connected in the new graph $\mathcal{G}'$.

A summary of the procedure is gathered in the Algorithm 1

Algorithm 1 Robust approximation of δ

Require: $\mathcal{E} = \{X_i \in \mathbb{R}^m,\ i = 1, \ldots, n\}$

Ensure: $\hat{\delta}$

1: Calculate $\mathrm{d}(X_i, X_{i'}) = \|X_i - X_{i'}\|_2$ between all pairwise data points X_i and $X_{i'}$, $i = 1, \ldots, n-1$, $i' = i+1, \ldots, n$, and construct the complete Euclidean graph $\mathcal{G}_{\mathrm{E}} = (\mathcal{E}, E)$ with set of edges $E = \{\{X_i, X_{i'}\}\}$.

2: Obtain the Euclidean Minimum Spanning Tree $\mathcal{G}_{\mathrm{MST}} = (\mathcal{E}, E_{\mathrm{T}})$ associated to $\mathcal{G}_{\mathrm{E}}$.

3: For each $i = 1, \ldots, n$ calculate $\epsilon_i = \max_{\{X_i, X_{i'}\} \in E_{\mathrm{T}}} \mathrm{d}\left(X_i, X_{i'}\right)$, and open balls $B\left(X_i, \epsilon_i\right) \subset \mathbb{R}^m$ of center X_i and radius ϵ_i. Construct a graph $\mathcal{G}' = (\mathcal{E}, E')$ adding more edges between points according to the rule

$$\{X_i, X_{i'}\} \in E' \Longleftrightarrow \overline{X_i X_{i'}} \subset \bigcup_{i=1}^{n} B\left(X_i, \epsilon_i\right),$$

where $\overline{X_i X_{i'}} = \{X \in \mathbb{R}^m,\ \exists \lambda \in [0,1],\ X = \lambda X_i + (1-\lambda) X_{i'}\}$.

4: Estimate the geodesic distance between two points by the length of the shortest path in the graph $\mathcal{G}'$ between these points using the Floyd's or Dijkstra's algorithm (see, e.g. Lee and Verleysen [23]).

Note that, the 3-step algorithm described above contains widespread graph-based methods to achieve our purpose. In this article, all graph-based calculations, such as Minimum Spanning Tree estimation or shortest path calculus, were carried out with the `igraph` package for network analysis by Csárdi and Nepusz [9].

An illustration of the algorithm and its behavior when the number of observations increases are displayed respectively in Figures 1 and 2. In Figure 1, points $\left(X_i^1, X_i^2\right)_i$ are simulated as follows:

$$X_i^1 = \frac{2i-n-1}{n-1} + \epsilon_i^1, \quad \text{and} \quad X_i^2 = 2\left(\frac{2i-n-1}{n-1}\right)^2 + \epsilon_i^2, \tag{4}$$

where ϵ_i^1 and ϵ_i^2 are independent and normally distributed with mean 0 and variance 0.01 for $i = 1, \ldots, n$ and $n = 30$. In Figure 2, some results of graph $\mathcal{G}'$ for

$n = 10, 30, 100$ are given. We can see that graph $\mathcal{G}'$ tends to be close to the true manifold $\left\{\left(t, t^2\right) \in \mathbb{R}^2,\ t \in \mathbb{R}\right\}$ when n increases.

Obviously, this estimation shows that the recovered structures in Figures 1 and 2 are pretty sensitive to noise. Nevertheless, to estimate a representative of a sample of curves, a prior smoothing step is almost always carried out as in Ramsay and Silverman [28]. This is done in Section 6 for our real data sets.

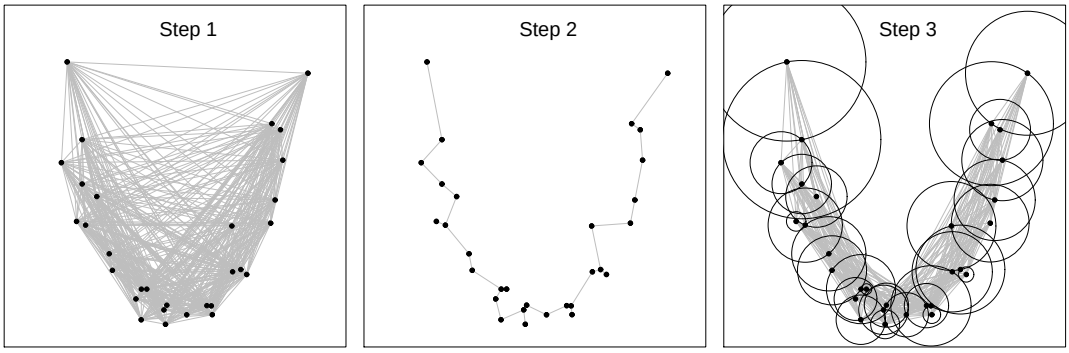

Figure 1: The 3-step construction of a subgraph $\mathcal{G}'$ from Simulation (4). On the left, the simulated data set (black dots) and the associated complete Euclidean graph $\mathcal{G}_{\mathrm{E}}$ (Step 1). On the middle, the $\mathcal{G}_{\mathrm{MST}}$ associated with the complete graph $\mathcal{G}_{\mathrm{E}}$ (Step 2). On the right, the associated open balls and the corresponding subgraph $\mathcal{G}'$ (Step 3).

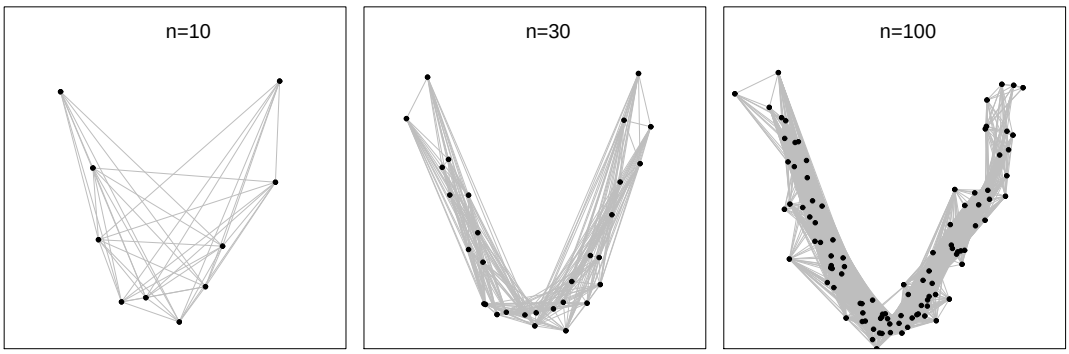

Figure 2: Evolution of graph $\mathcal{G}'$ from Simulation (4) for $n = 10, 30, 100$.

4 Application: Template estimation in a shape invariant model

In this section, we consider the case where the observations are curves warped from an unknown template $f : \mathcal{T} \to \mathbb{R}$. We want to study whether the *central* curve defined previously as the median of the data with respect to the geodesic distance provides a good pattern of the curves. Good means, in that particular case, that the intrinsic median should be close to the pattern f.

We consider a translation model indexed by a real valued random variable A with unknown distribution on an interval $(b, c) \subset \mathbb{R}$

$$X_{ij} = f_i(t_j) = f\left(t_j - A_i\right), \; i = 1, \ldots, n, \; j = 1, \ldots, m, \tag{5}$$

where $(A_i)_i$ are i.i.d random variables drawn with distribution A which models the unknown shift parameters. This specification is an special case of the self-modeling regression mentioned in the introduction.

Under a nonparametric registration model, Maza [25] and Dupuy et al. [14] define the *structural expectation* function of a sample of curves and build a registration procedure in order to estimate it. Following the same philosophy, but for the case of the translation effect model (5), we propose to use as a good pattern of the dataset the *Structural Median* function f_{SM} defined as

$$f_{\mathrm{SM}} = f\left(\cdot - \mathrm{med}(A)\right), \tag{6}$$

where $\mathrm{med}(A)$ denotes the median of A.

We will see that the manifold embedding point of view enables to recover this pattern. Actually, define a one-dimensional function in $\mathcal{M} \subset \mathbb{R}^m$ parameterized by a parameter $a \in (b, c) \subset \mathbb{R}$ as

$$\begin{aligned} X \colon (b, c) &\to \mathbb{R}^m \\ a \quad &\mapsto X(a) = \left(f\left(t_1 - a\right), \ldots, f\left(t_m - a\right)\right), \end{aligned}$$

and set $\mathcal{C} = \{X(a) \in \mathbb{R}^m, \; a \in (b, c)\}$.

As soon as X is a regular curve, that is, if its first derivative never vanishes,

$$X' \neq 0 \Longleftrightarrow \forall a \in (b, c), \; \exists j \in \{1, \ldots, m\}, \; f'\left(t_j - a\right) \neq 0, \tag{7}$$

then, the smooth mapping $X \colon a \mapsto X(a)$ provides a natural parametrization of $\mathcal{C}$ which can thus be seen as a submanifold of $\mathbb{R}^m$ of dimension 1 (do Carmo [12]). The corresponding geodesic distance is given by

$$\delta\left(X(a_1), X(a_2)\right) = \left| \int_{a_1}^{a_2} \left\| X'(a) \right\| \mathrm{d}a \right|, \tag{8}$$

with $X'(a) = \mathrm{d}X(a)/\mathrm{d}a = \left(\mathrm{d}X_1(a)/\mathrm{d}a, \ldots, \mathrm{d}X_m(a)/\mathrm{d}a\right)^\top$.

The observation model (5) can then be seen as a discretization of the manifold $\mathcal{C}$ for different values $(A_i)_i$. Hence, finding the intrinsic median of all shifted curves can be done by understanding the geometry of space $\mathcal{C}$, and thus, by approximating the geodesic distance between observed curves. Define the intrinsic median with respect to the geodesic distance (8) on $\mathcal{C}$, that is

$$\widehat{\mu}_I^1 = \arg\min_{\mu \in \mathcal{C}} \sum_{i=1}^{n} \delta\left(X_i, \mu\right). \tag{9}$$

The following theorem gives a minimizer.

Theorem 1. *Under the assumption (7) that X is a regular curve, we get that*

$$\widehat{\mu}_I^1 = \left(f\left(t_1 - \widehat{\mathrm{med}}(A)\right), \ldots, f\left(t_m - \widehat{\mathrm{med}}(A)\right)\right),$$

where $\widehat{\mathrm{med}}(A)$ is the empirical median.

Hence as soon as we observe a sufficient number of curves to ensure that the median and the empirical median are close, the intrinsic median is a natural approximation of (6). Therefore, the manifold framework provides a geometrical interpretation of the structural median of a sample of curves. The estimator is thus given by

$$\widehat{\mu}_{I,n}^1 = \arg\min_{\mu \in \mathcal{E}} \sum_{i=1}^{n} \widehat{\delta}\left(X_i, \mu\right), \tag{10}$$

where $\widehat{\delta}$ is an approximation of the unknown underlying geodesic distance, that is estimated by the algorithm described in Section 3.

We point out that in many situations, giving a particular model for the deformations corresponds actually to consider a particular manifold embedding for the data. Once the manifold is known, its corresponding geodesic distance may be properly computed, as done in the translation case. So in some particular cases, the minimization in (9) can give an explicit formulation and then it is possible to identify the resulting Fréchet median. Hence previous theorem may be generalized to such cases as done in Gallón et al. [15].

Note first that this case only holds for the Fréchet median ($\alpha = 1$) but not the mean for which the so-called structural expectation and the Fréchet mean are different. Moreover, the choice of the median is also driven by the need for a robust method, whose good behavior will be highlighted in the simulations and applications in the following sections.

As shown in the simulation study below, when the observations can be modeled by a set of curves warped from an unknown template by a general deformation process, estimate (10) enables to recover the main pattern in a better way than classical methods. Obviously, the method relies on the assumption that all the observed data belong to an embedded manifold whose geodesic distance can be well approximated by the proposed algorithm.

5 Simulation study

In this section, the numerical properties of our estimator, called Robust Manifold Embedding (RME), defined by the equation (10) in Section 4 are studied using simulated data. The estimator is compared to those obtained with the Isomap algorithm and the Modified Band Median (MBM) estimator proposed by Arribas-Gil and Romo [1], which is based on the concept of depth for functional data (see López-Pintado and Romo [24]). The behavior of the estimator when the number of curves is increased is also analyzed.

Four different types of simulations of increasing warping complexity for the single shape invariant model were carried out, observing $n = 15, 30, 45, 60$ curves on $m = 100$ equally spaced discrete points $(t_j)_j$ in the interval $[-10, 10]$. The experiment was conducted with $R = 100$ repetitions. The template function f and shape invariant model, for each simulation, are given as follows:

***Simulation 1*:** One-dimensional manifold defined by $f(t) = 5\sin(t)/t$ and

$$X_{ij} = f\left(t_j + A_i\right),$$

where $(A_i)_i$ are i.i.d uniform random variables on interval $[-5, 5]$.

***Simulation 2*:** Two-dimensional manifold given by $f(t) = 5\sin(t)$ and

$$X_{ij} = f\left(A_i t_j + B_i\right),$$

where $(A_i)_i$ and $(B_i)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[0.7, 1.3]$ and $[-1, 1]$.

***Simulation 3*:** Four-dimensional manifold given by $f(t) = t\sin(t)$ and

$$X_{ij} = A_i f\left(B_i t_j + C_i\right) + D_i,$$

where $(A_i)_i$, $(B_i)_i$, $(C_i)_i$ and $(D_i)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[0.7, 1.3]$, $[0.7, 1.3]$, $[-1, 1]$ and $[-1, 1]$.

***Simulation 4*:** Four-dimensional manifold given by $f(t) = \phi t + t\sin(t)\cos(t)$ with $\phi = 0.9$, and

$$X_{ij} = A_i f\left(B_i t_j + C_i\right) + D_i,$$

where $(A_i)_i$, $(B_i)_i$, $(C_i)_i$ and $(D_i)_i$ as in the Simulation 3.

Figure 3 illustrates the simulated data sets from Simulations 1-4 with $n = 30$ curves for one simulation run (one of 100 repetitions). For Simulation 1, where there is only phase variability, all methods follows the structural characteristics of the sample of curves, where the template estimated by the robust manifold approach is the closest curve to the theoretical function. The same conclusion can be inferred from Simulation 2. Indeed, for this simulation type, and for this particular simulation run, the RME estimator coincides with the theoretical template function. For the four-dimensional manifolds in Simulations 3 and 4, where there is an additional amplitude variability, the robust manifold estimator captures better the structural pattern in the sample of curves followed by the Isomap estimator. Note that in the Simulation 4, both approaches coincide. Although the MBM estimator follows the shape of the theoretical template, the estimator deviates from it in the cases 2-4.

In order to compare more accurately the estimators described above, we calculate, for each one, the empirical mean squared error obtained on the $R = 100$ repetitions of each type of simulation. We recall the definition, for estimator $\hat{f}$ of a given type of simulation, of the mean squared error:

$$\text{Mean Squared Error}\left(\hat{f}\right) = \frac{1}{R}\sum_{r=1}^{R} \|\hat{f}_r - f\|_2^2,$$

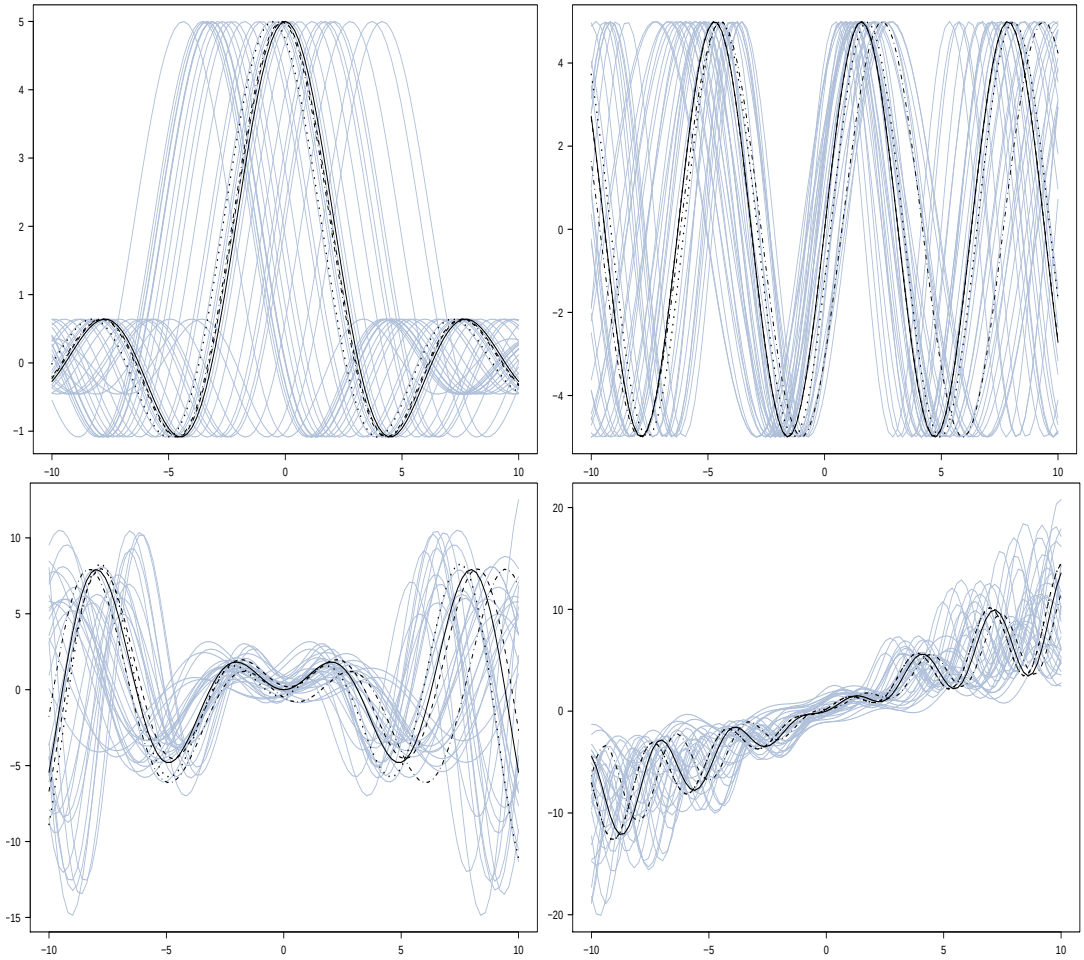

Figure 3: Simulated curves (gray) from Simulation 1 (top left), 2 (top right), 3 (bottom left) and 4 (bottom right) for one simulation run, and the respective target template function f (solid line), MBM (dash-dotted line), Isomap (dotted line), and RME (dashed line) template estimators.

where, $\hat{f}_r$ is the estimation from the r-th repetition of the given simulation type, f is the true template function and $\|\cdot\|_2$ is the classical Euclidean norm. We also highlight, for our comparisons, the fact that

$$\text{Mean Squared Error}\left(\hat{f}\right) = \underbrace{\frac{1}{R}\sum_{r=1}^{R}\|\hat{f}_r - \bar{f}\|_2^2}_{\text{Variance}} + \underbrace{\|\bar{f} - f\|_2^2}_{\text{Squared bias}},$$

where $\bar{f}$ is the mean of all R obtained estimations.

Table 1 shows the mean squared errors, variances and squared biases of each estimator for simulations 1, 2, 3 and 4, and for different number on curves $n = 15, 30, 45, 60$ in the sample. Values have been rounded to zero decimal places to facilitate the comparisons, and the minimum values are signed in bold.

From the table, we can observe that when the number of curves in the sample is small ($n = 15$) the MBM estimator has better results in terms of the MSE, Bias2 and variance for the Simulation 1. The same is true when $n = 30$. With $n = 45, 60$ curves the MBM estimator has minimum mean squared error and variance, and our estimator has smaller bias. Comparing the Isomap and RME methods only, the latter overcomes the former. For Simulation 2, the RME estimator overcomes the MBM and Isomap estimators for $n = 30, 45, 60$ curves, except when $n = 15$, where the Isomap estimator is better. However, in this case there are not big differences between Isomap and RME estimators. As we expected, when the geometry of the curves is more complex, i.e. when we have a four-dimensional manifolds, the results are more variated. For Simulation 3, the RME estimator has a good performance with $n = 45, 60$. With $n = 15, 30$ the better results are shared by the MBM and Isomap methods. In Simulation 4, the MBM estimator has, in general, better results. Finally, note that although the theorem in Section 4 is valid for one-dimensional manifolds generated by time shifts (Simulation 1), we can see that the intrinsic sample median estimator by approximating the corresponding geodesic distance with the robust algorithm performs well for manifolds of higher dimension (Simulations 2-4).

Table 1: Comparison of estimators for Simulations 1-4 with different sample sizes.

n	Statistic	Simulation1			Simulation 2		
		MBM	Isomap	RME	MBM	Isomap	RME
15	MSE	**136**	389	335	790	**400**	435
	Bias2	**23**	152	118	141	**35**	46
	Variance	**113**	236	217	649	**366**	389
30	MSE	**30**	108	92	666	338	**268**
	Bias2	**8**	13	10	98	34	**18**
	Variance	**22**	95	82	568	304	**249**
45	MSE	**24**	139	66	669	244	**155**
	Bias2	10	23	**5**	120	20	**9**
	Variance	**14**	116	60	549	224	**147**
60	MSE	**14**	85	55	634	168	**136**
	Bias2	5	9	**4**	161	5	**4**
	Variance	**10**	76	51	472	163	**132**
n	Statistic	Simulation 3			Simulation 4		
		MBM	Isomap	RME	MBM	Isomap	RME
15	MSE	1350	**1152**	1171	**876**	890	893
	Bias2	**251**	375	441	**394**	512	522
	Variance	1098	777	**730**	483	378	**371**
30	MSE	1025	**673**	721	911	**861**	876
	Bias2	212	**160**	248	**470**	536	554
	Variance	813	513	**473**	441	325	**323**
45	MSE	1034	553	**498**	**820**	827	868
	Bias2	223	**105**	141	**397**	524	569
	Variance	811	449	**356**	423	303	**299**
60	MSE	965	572	**402**	879	**776**	842
	Bias2	168	**97**	122	**458**	474	563
	Variance	797	475	**280**	421	302	**279**

Robustness analysis

In order to assess the robustness of the RME estimator, we carried out an additional simulation study generating atypical curves in the functional data sets. Atypical curves can be considered in the two following settings: either real outliers which do not share any common shape with the observations or curves that are obtained by atypical deformations within the same model. The methodology that we propose is based on the existence of a common shape between the data, hence adding complete outliers may break the geometry of the data, i.e. the manifold structure, if the outliers are far from any deformations. This remark is true for any registration method when curves away from the warping model are considered. Hence robust methods without the constraint of a deformation model, for instance based on the notion of depth function, behave better. Yet, we show in the following that when considering curves warped by a deformation process which has some outliers, our method is robust with respect to the Isomap procedure and may compete with the MBM estimator.

Thereby, from $n = 15, 30, 45, 60$ curves we generated 10% of them as atypical according to the single shape invariant model in the four type of simulations considered above, modifying the corresponding individual shift parameters but preserving the geometric structure of the curves. So, for each simulation, the non-atypical curves X_{ij} with $i = 1, \ldots, (n - \lceil 0.10n \rceil)$ are generated as above, and the atypical curves $\tilde{X}_{ij}$ with $i = (n - \lceil 0.10n \rceil) + 1, \ldots, n$ were generated as:

***Simulation 1*:** One-dimensional manifold defined by $f(t) = 5\sin(t)/t$ and

$$\tilde{X}_{ij} = f\left(t_j + \tilde{A}_i\right),$$

where $\left(\tilde{A}_i\right)_i$ are i.i.d uniform random variables on interval $[4.5, 6]$.

***Simulation 2*:** Two-dimensional manifold given by $f(t) = 5\sin(t)$ and

$$\tilde{X}_{ij} = f\left(\tilde{A}_i t_j + \tilde{B}_i\right),$$

where $\left(\tilde{A}_i\right)_i$ and $\left(\tilde{B}_i\right)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[0.35, 0.65]$ and $[-0.5, 0.5]$.

***Simulation 3*:** Four-dimensional manifold given by $f(t) = t\sin(t)$ and

$$\tilde{X}_{ij} = \tilde{A}_i f\left(\tilde{B}_i t_j + \tilde{C}_i\right) + \tilde{D}_i,$$

where $\left(\tilde{A}_i\right)_i$, $\left(\tilde{B}_i\right)_i$, $\left(\tilde{C}_i\right)_i$ and $\left(\tilde{D}_i\right)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[1.3, 1.4]$, $[0.7, 1.3]$, $[-1.5, -1]$ and $[1, 1.5]$.

***Simulation 4*:** Four-dimensional manifold given by $f(t) = \phi t + t\sin(t)\cos(t)$ with $\phi = 0.9$, and

$$\tilde{X}_{ij} = \tilde{A}_i f\left(\tilde{B}_i t_j + \tilde{C}_i\right) + \tilde{D}_i,$$

where $\left(\tilde{A}_i\right)_i$, $\left(\tilde{B}_i\right)_i$, $\left(\tilde{C}_i\right)_i$ and $\left(\tilde{D}_i\right)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[1.05, 1.95]$, $[1.05, 1.95]$, $[-1, 1]$ and $[-1, 1]$.

Figure 4 illustrates the simulated data sets from Simulations 1-4 with $n = 30$ curves for one replication. The curves signed in red color correspond to the atypical curves. For this particular simulation run, we see that the atypical curves has influence over the Isomap estimator for all types of simulation. For the one and two-dimensional manifolds in Simulations 1 and 2 respectively we observe that the RME estimator has a good performance. For example, note, as in the simulation study without atypical curves developed above, the RME estimator coincides with the theoretical template function for the simulation type 1, and for this particular simulation run. For complex shape functions as in Simulations 3 and 4, our estimator captures adequately the common pattern of the sample in presence on atypical curves. As expected, the depth-based estimator is robust against atypical observations.

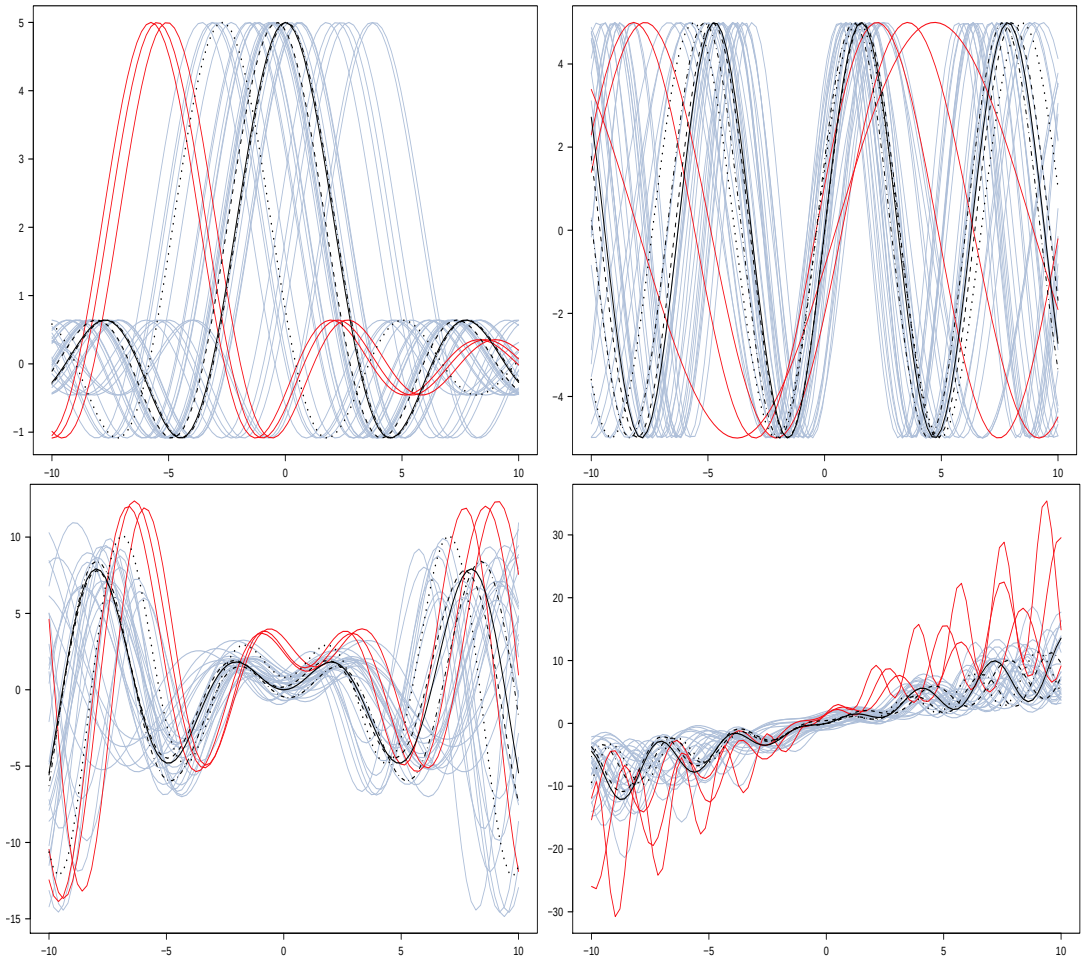

Figure 4: Simulated curves (gray) from simulation type 1 (top left), 2 (top right), 3 (bottom left) and 4 (bottom right) including atypical curves (red) for one simulation run, and the respective target template function f (solid line), MBM (dash-dotted line), Isomap (dotted line), and RME (dashed line) template estimators.

The mean squared errors, variances and squared biases of each estimator for Simulations 1-4 and different number on curves $n = 15, 30, 45, 60$ including atypical

curves are showed in the Table 2. For Simulation 1, the MBM estimator has the best results for all number of curves. In this case, the RME method overcomes its not robust version estimator (Isomap). Additionally, when the warping complexity increases, the RME estimator has minimum mean squared errors in most cases for Simulations 2-4. As expected, only the when the number of curves is small ($n = 15$) the estimator performs less well.

Table 2: Comparison of estimators including atypical curves for different sample sizes.

n	Statistic	Simulation1			Simulation 2		
		MBM	Isomap	RME	MBM	Isomap	RME
15	MSE	**107**	452	400	830	**570**	680
	Bias2	**21**	215	169	186	**76**	116
	Variance	**85**	237	232	645	**495**	564
30	MSE	**36**	177	166	649	409	**300**
	Bias2	**8**	49	46	115	39	**20**
	Variance	**28**	128	120	535	370	**280**
45	MSE	**21**	121	81	523	296	**212**
	Bias2	**9**	32	26	89	23	**11**
	Variance	**13**	89	56	433	273	**200**
60	MSE	**19**	151	90	551	276	**212**
	Bias2	**9**	45	40	98	63	**46**
	Variance	**10**	106	51	453	213	**166**
n	Statistic	Simulation 3			Simulation 4		
		MBM	Isomap	RME	MBM	Isomap	RME
15	MSE	1387	**1093**	1098	990	983	**963**
	Bias2	**257**	327	396	**485**	561	538
	Variance	1129	766	**702**	505	**422**	425
30	MSE	1370	**856**	857	901	861	**856**
	Bias2	260	**204**	312	**434**	505	511
	Variance	1110	652	**545**	467	355	**345**
45	MSE	1206	640	**547**	874	863	**860**
	Bias2	234	**155**	197	556	541	**406**
	Variance	972	484	**350**	**317**	322	454
60	MSE	963	585	**462**	924	864	**861**
	Bias2	154	**118**	165	**500**	537	561
	Variance	809	468	**297**	424	327	**301**

6 Applications

In this section we apply the proposed robust manifold learning algorithm to extract the template function of a sample of curves on three real datasets of functional data: the well-known Berkeley Growth and Gait data in functional data applications (Ramsay and Silverman [28]), and a reflectance data of two landscape types. Our algorithm is compared with the Isomap and Modified Band Median methods.

6.1 Berkeley growth study

The data of the Berkeley's study consist in 31 height measurements for 54 girls and 38 boys recorded between the ages of 1 and 18 years. Intervals between measurements range from 3 months (age 1-2 years), to yearly (age 3-8), to half-yearly (age 8-18). One of the goals with this kind of data is the pattern analysis of growth velocity and acceleration curves, represented by the first and second derivatives of the height functions, in order to characterize its spurts and trends during years. The velocity and acceleration curves for girls and boys were obtained by taking the first and second order differences, respectively, of the height curves, whose functional representations were made using a B-spline smoothing (see Ramsay and Silverman [28] for details).

Figure 5 provides the smoothed velocity curves (on the top) measured in centimeters per year (cm/year) and the smoothed acceleration curves (on the bottom) measured in centimeters per squared year (cm/year^2) of height for girls (on the left) and for boys (on the right). It is evident that all individuals exhibit a common velocity and acceleration pattern throughout years, but features as peaks, troughs and inflection points occur at different times for each child.

From all of the graphs in the Figure 5, we see, in general, that all the template estimators obtain a curve situated in the middle of the samples of curves capturing its common shape pattern appropriately. For the case of sample velocity curves of girls (top-left graph) the RME and MBM estimators coincide. The Isomap estimator deviates slightly from the center. In the case of samples of velocity and acceleration curves of boys, both the RME and Isomap estimators choose the same template function. Only in the case of acceleration curves of girls (bottom-left graph), the three methods choose different functions. In summary, we infer that the RME estimator seems to perform a good work extracting a meaningful shape curve.

6.2 Gait cycle data

For this application, we consider the data of angle measurements (in degrees) in the sagittal plane formed by the hip and knee of 39 children through a gait cycle, where time is measured in terms of the child's gait cycle such that every curve is given for values ranging between 0 and 1. The smoothed curves were obtained by fitting a Fourier basis system following the analysis of Ramsay and Silverman [28] for this data, where both sets of curves are periodic. Figure 6 displays the curves of hip (on the left) and knee (on the right) angles observed during the gait. As we can see, a two-phase process can be identified for the knee motion, while for the hip motion there is a single-phase. Also, both sets of curves share a common pattern around which there are both phase and amplitude variability.

For this application, the template functions obtained by the Robust Manifold Estimator based on our algorithm seem to capture the salient features of the sample of hip and knee angle curves. Note also that the same template, located in the center of the samples, was chosen by all the estimators.

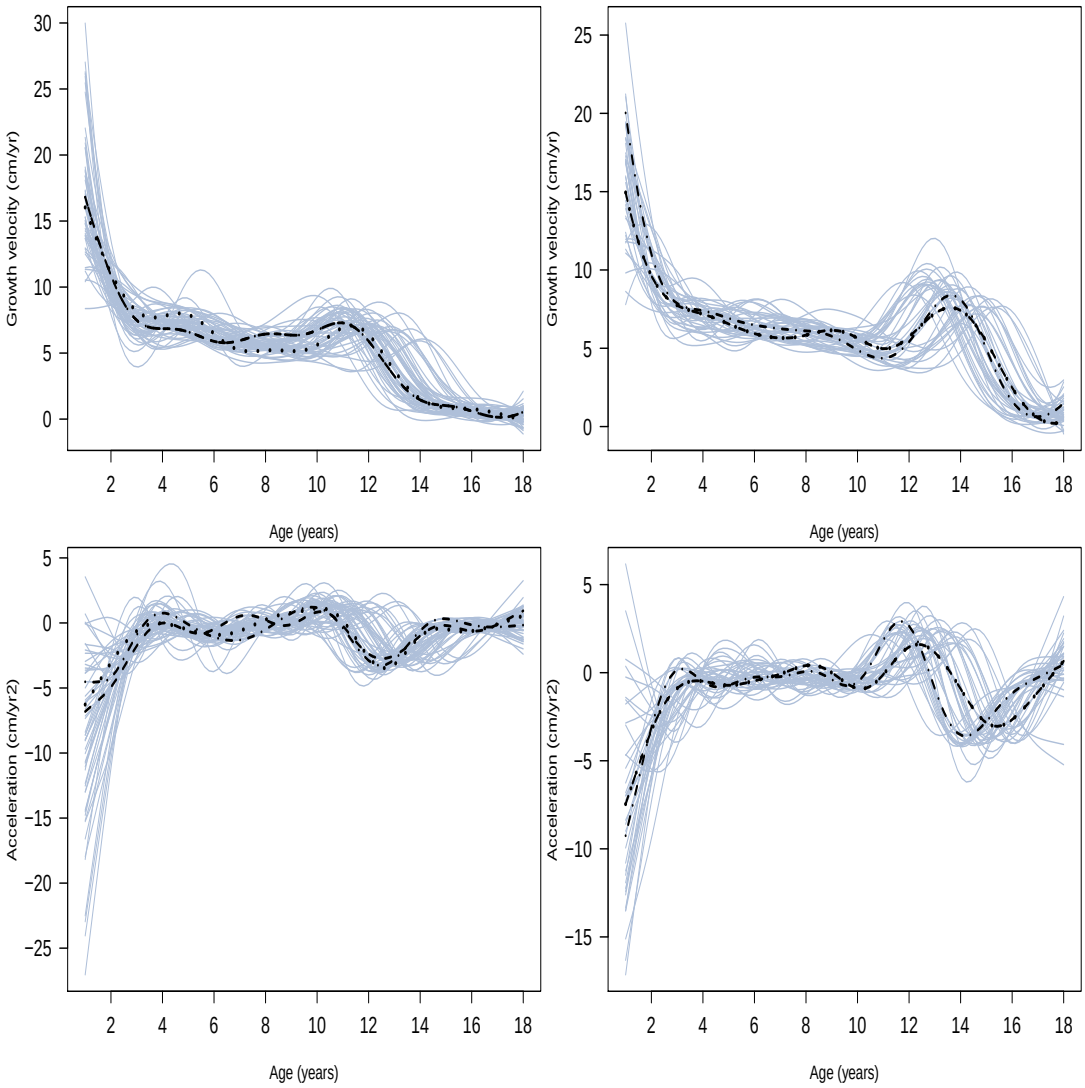

Figure 5: Velocity (on the top) and acceleration (on the bottom) curves of 54 girls (on the left) and 31 boys (on the right) in the Berkeley growth study (gray lines). The estimated template functions with the MBM (dashed line), Isomap (dashed-dotted line), and RME (solid line) methods.

6.3 Landscape reflectances data

Finally, we consider two data sets where the corresponding observed curves represent the weekly reflectance profiles of two particular landscapes (corn and wheat). The reflectance is a measure of the incident electromagnetic radiation that is reflected by a given interface. For these data, there are 23 and 124 curves for corn and wheat landscapes respectively. The aim consists in extracting a representative curve of a type of landscape while observing the reflectance profiles of different landscapes of the same type. In Figure 7, the smoothed curves corresponding to reflectance patterns of two

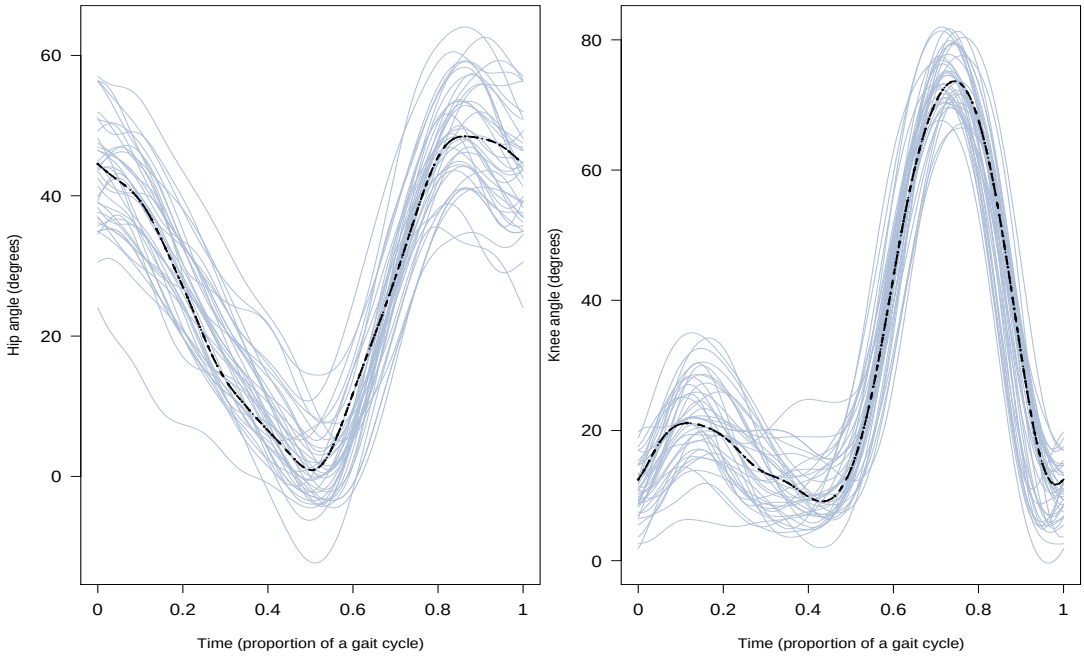

Figure 6: Angle curves formed by the hip (on the left) and knee (on the right) of 39 children through a gait cycle, and the MBM (dash-dotted line), Isomap (dotted line), and RME (dashed line) template estimators.

landscape types in the same region in the same period are showed. The smoothing was obtained from discrete data with B-spline basis system. The reflectance depends on the vegetation whose growth depends on the weather condition and the soil behavior. It is therefore relevant to consider that these profiles are deformations in translation, scale and amplitude of a single representative function of the reflectance behavior of each landscape type in this region at this time.

In Figure 7, we observe that for the corn landscape case, where there are relatively a few number of curves, the robust manifold estimator chooses a meaningful template curve which seems to appear at the center of curve sample, which coincides with the curve obtained by the modified band median estimator. The Isomap estimator chooses a different curve as representative function which is slightly away from the center of the sample. For the wheat landscape, all of three estimators choose a different template curve. Although all the estimated template curves follow the structural features of the sample, the RME estimator select a curve that is located more in the middle. In this application domain, extracting a curve by RME is best able to report data as reflecting their structure and thus to obtain a better representative and improve further future functional analysis.

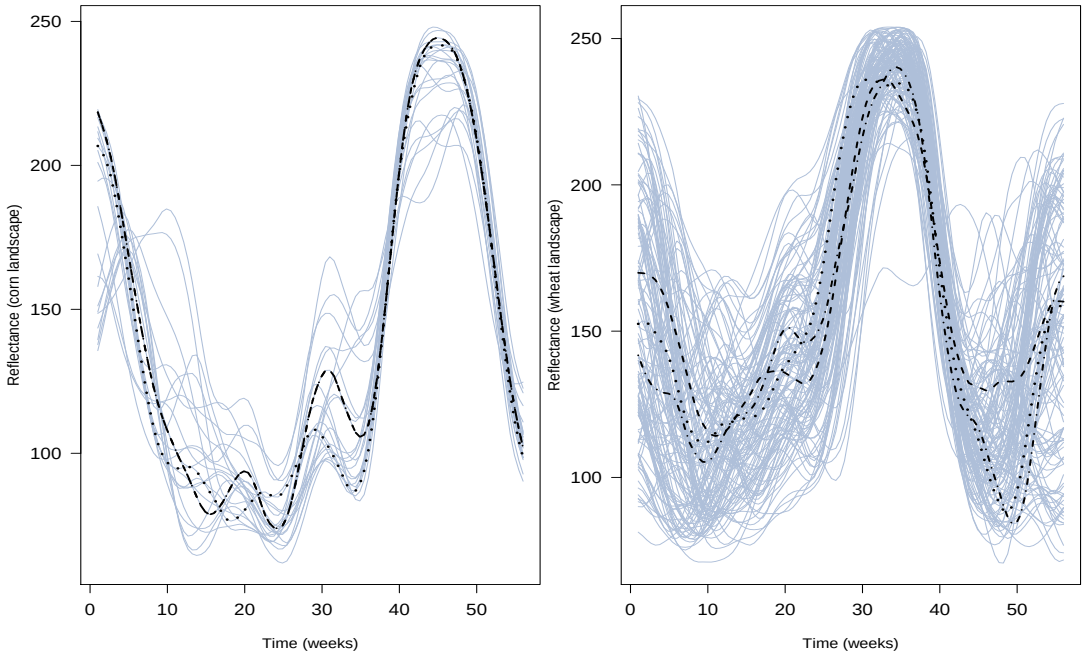

Figure 7: Reflectance curves of corn (left) and wheat (right) landscapes, and MBM (dash-dotted line), Isomap (dotted line), and RME (dashed line) template estimators.

7 Concluding remarks

In this paper, we have proposed a robust algorithm to approximate the geodesic distance of the underlying manifold. This approximated distance is used to build the corresponding empirical Fréchet median of the functions. This function is a meaningful template curve for a sample of functions, which have both amplitude and time deformations.

Our approach relies on the fundamental paradigm of functional data analysis which involves treating the entire observed curve as a unit of observation rather than individual measurements from the curve. Indeed, we show that, when the structure of the deformations entails that the curve can be embedded into a manifold, finding a representative of a sample of curves corresponds to calculate an intrinsic statistic of observed curves on their unknown underlying manifold. Moreover in a translation model, i.e. where the curves are actually warped from an unknown pattern, both methodologies coincide since the structural median of a sample of curves corresponds to the intrinsic median on a one-dimensional manifold. Moreover, our methodology could be the first step for classification methods which require the choice of a good template, sharing the common properties of the dataset, see for instance in Shimizu and Mizuta [32], Cuesta-Albertos and Fraiman [10], Sangalli et al. [31] and references therein.

From a computational point of view, our method is inspired by the ideas of the Isomap algorithm. We note that we have also used the Isomap algorithm in the

simulation study and applications. Hence, our algorithm has the advantage of being parameter free and then it is of easiest use. One of the major drawbacks of these methodologies are that a relatively high number of data are required in order to guarantee a good approximation of the geodesic distance at the core of this work (see Tenenbaum et al. [35]). Nevertheless, our method is a robust version of the Isomap and behaves better with respect to it. In addition it competes with a robust algorithm of the modified band median estimator. The `R` code is available at the webpage of one of the authors (`http://santiagogallongomez.wordpress.com/code/`) or upon request.

8 Appendix

Proof of Theorem 1. Let X be defined by

$$\begin{aligned} X\colon (b,c) &\to \mathbb{R}^m \\ a &\mapsto X(a) = \left(f\left(t_1 - a\right), \ldots, f\left(t_m - a\right)\right) \end{aligned}$$

and set $\mathcal{C} = \{X(a) \in \mathbb{R}^m,\ a \in (b,c)\}$.

By assumption (7), $\mathcal{C}$ can be seen as a submanifold of $\mathbb{R}^m$ of dimension 1 with corresponding geodesic distance defined by (8).

Take $\mu = X(\alpha)$ with $\alpha \in (b,c)$, thus we can write

$$\begin{aligned} \widehat{\mu}_I^1 &= \underset{X(\alpha)\in\mathcal{C}}{\arg\min} \sum_{i=1}^{n} \delta\left(X\left(A_i\right), X(\alpha)\right) \\ &= \underset{\mu\in\mathcal{C}}{\arg\min} \sum_{i=1}^{n} D\left(A_i, \alpha\right) = \underset{\mu\in\mathcal{C}}{\arg\min}\, C(\alpha) \end{aligned}$$

where D is distance on (b,c) given by

$$D\left(A_i, \alpha\right) = \left| \int_{A_i}^{\alpha} \left\| X'(a) \right\| \mathrm{d}a \right|.$$

In the following, let $\left(A_{(i)}\right)_i$ be the ordered parameters. That is $A_{(1)} < \cdots < A_{(n)}$. Then, for a given $\alpha \in (b,c)$ such that $A_{(j)} < \alpha < A_{(j+1)}$, we get that

$$\begin{aligned} C(\alpha) = {} & jD\left(\alpha, A_{(j)}\right) + \sum_{i=1}^{j-1} iD\left(A_{(i)}, A_{(i+1)}\right) \\ & + (n-j)D\left(\alpha, A_{(j+1)}\right) + \sum_{i=j+1}^{n-1} (n-i)D\left(A_{(i)}, A_{(i+1)}\right). \end{aligned}$$

For the sake of simplicity, let $n = 2q+1$. It follows that $\widehat{\mathrm{med}}(A) = A_{(q+1)}$. Moreover, let $\alpha = A_{(j)}$ with $j < q+1$. By symmetry, the case $j > q+1$ holds. Then, we rewrite $C\left(\alpha\right)$ as

$$C\left(\alpha\right) = \sum_{i=1}^{j-1} iD\left(A_{(i)}, A_{(i+1)}\right) + \sum_{i=j}^{n-1} (n-i)D\left(A_{(i)}, A_{(i+1)}\right)$$

and, by introducing $A_{(q+1)}$, we get that

$$C(\alpha) = \sum_{i=1}^{j-1} iD\left(A_{(i)}, A_{(i+1)}\right) + \sum_{i=j}^{q} iD\left(A_{(i)}, A_{(i+1)}\right)$$
$$+ \sum_{i=j}^{q}(n-2i)D\left(A_{(i)}, A_{(i+1)}\right) + \sum_{i=q+1}^{n-1}(n-i)D\left(A_{(i)}, A_{(i+1)}\right).$$

Finally, we notice that

$$C(\alpha) = C\left(A_{(q+1)}\right) + \sum_{i=j}^{q}(n-2i)D\left(A_{(i)}, A_{(i+1)}\right) > C\left(A_{(q+1)}\right).$$

And the result follows since

$$\widehat{\mu}_I^1 = \arg\min_{\mu\in\mathcal{C}} C(\alpha) = X\left(A_{(q+1)}\right) = X\left(\widehat{\mathrm{med}}(A)\right) = \widehat{f}_{\mathrm{SM}}.$$

■

References

[1] A. Arribas-Gil and J. Romo. Robust depth-based estimation in the time warping model. *Biostatistics*, 13:398–414, 2012.

[2] M. Balasubramanian and E. L. Schwartz. The isomap algorithm and topological stability. *Science*, 295, 2002.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15:1373–1396, 2003.

[4] M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University, 2000. Available at http://isomap.stanford.edu/BdSLT.pdf.

[5] J. Bigot, J.-M. Loubes, and M. Vimond. Semiparametric estimation of shifts on compact Lie groups for image registration. *Probab. Theory Related Fields*, 152: 425–473, 2012.

[6] I. Castillo and J.-M. Loubes. Estimation of the distribution of random shifts deformation. *Math. Methods Statist.*, 18:21–42, 2009.

[7] D. Chen and H. G. Müller. Nonlinear manifold representations for functional data. *Ann. Statist.*, 40:1–29, 2012.

[8] R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21: 5–30, 2006.

[9] G. Csárdi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. Available at http://igraph.sf.net.

[10] J. Cuesta-Albertos and R. Fraiman. Impartial trimmed k-means for functional data. *Comput. Statist. Data Anal.*, 51:4864–4877, 2007.

[11] V. de Silva and J. B. Tenenbaum. Unsupervised learning of curved manifolds. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear estimation and classification*, volume 171 of *Lecture Notes in Statistics*, pages 453–465. Springer-Verlag, New York, 2003.

[12] M. P. do Carmo. *Riemannian Geometry*. Mathematics: Theory & Applications. Birkhäuser Boston Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.

[13] D. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA*, 100:5591–5596, 2003.

[14] J. F. Dupuy, J. M. Loubes, and E. Maza. Non parametric estimation of the estructural expectation of a stochastic increasing function. *Stat. Comput.*, 21: 121–136, 2011.

[15] S. Gallón, J.-M. Loubes, and E. Maza. Statistical properties of the quantile normalization method for density curve alignment. *Math. Biosci.*, 2013. http://dx.doi.org/10.1016/j.mbs.2012.12.007.

[16] F. Gamboa, J. Loubes, and E. Maza. Semi-parametric estimation of shits. *Electron. J. Stat.*, 1:616–640, 2007.

[17] G. James. Curve alignment by moments. *Ann. Appl. Stat.*, 1:480–501, 2007.

[18] A. Kneip and T. Gasser. Convergence and consistency results for self-modelling regression. *Ann. Statist.*, 16:82–112, 1988.

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

[20] A. Kneip and J. Ramsay. Combining registration and fitting for funstional models. *J. Amer. Statist. Assoc.*, 103:1155–1165, 2008.

[21] A. Kneip, X. Li, X. MacGibbon, and J. Ramsay. Curve registration by local regression. *Canad. J. Statist.*, 28:19–29, 2000.

[22] R. Koenker. The median is the message: Toward the Fréchet median. *J. Soc. Fr. Stat.*, 147:61–64, 2006.

[23] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer, New York, 2007.

[24] S. López-Pintado and J. Romo. On the concept of depth for functional data. *J. Amer. Statist. Assoc.*, 104:718–734, 2009.

[25] E. Maza. Estimation de l'espérance structurelle d'une fonction aléatoire. *C. R. Math. Acad. Sci. Paris*, 343:551–554, 2006.

[26] X. Pennec. Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. *J. Math. Imaging Vision*, 25:127–154, 2006.

[27] J. O. Ramsay and X. Li. Curve registration. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60:351–363, 1998.

[28] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, 2nd edition, 2005.

[29] B. Rønn. Nonparametric maximum likelihood estimation for shifted curves. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63:243–259, 2001.

[30] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear

embedding. *Science*, 290:2323–2326, 2000.

[31] L. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. k-mean alignment for curve clustering. *Comput. Statist. Data Anal.*, 54:1219–1233, 2010.

[32] N. Shimizu and M. Mizuta. Functional clustering and functional principal points. In B. Apolloni, R. J. Howlett, and L. Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 4693 of *Lecture Notes in Computer Science*, pages 501–508. Springer, Berlin Heidelberg, 2007.

[33] B. W. Silverman. Incorporating parametric effects into functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57:673–689, 1995.

[34] R. Tang and H.-G. Müller. Pairwise curve synchronization for functional data. *Biometrika*, 95:875–889, 2008.

[35] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[36] T. Trigano, U. Isserles, and Y. Ritov. Semiparametric curve alignment and shift density estimation for biological data. *IEEE Trans. Signal Process.*, 59:1970–1984, 2011.

[37] S. Vantini. On the definition of phase and amplitude variability in functional data analysis. *Test*, 21:676–696, 2012.

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

[39] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.*, 26:313–338, 2004.