A robust algorithm for template curve estimation based on manifold embedding

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

A robust algorithm for template curve estimation based on manifold embedding

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

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

[b] *Departamento de Matemáticas y Estadística, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia.*

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

February 8, 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 $\{f_i(t) : t \in \mathcal{T} \subset \mathbb{R}, i = 1, \ldots, n\}$ 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

[*] cd@geosys.com

[†] santiagog@udea.edu.co

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

[§] Elie.Maza@ensat.fr

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 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 [24], Ramsay and Silverman [25], or Vantini [32] 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 [16]).

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 the 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 [16], Wang and Gasser [33] Kneip et al. [18], James [14], Tang and Müller [29], and Kneip and Ramsay [17] or Dupuy et al. [11]. 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 [15]), where all functions are deduced with respect to the template f by mean an individual parameter vector θ_i. This point of view is also followed in Silverman [28], Rønn [26], Gamboa et al. [13], Castillo and Loubes [6], Bigot et al. [5] and Trigano et al. [31].

The second category of methods do not assume any deformation model for the individual functions. The purpose is to select a curve 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 [21] 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 the mean of the data with respect to this geodesic distance, hence approximating the Fréchet mean or median of the data. Several algorithms have been developed over the last decade in order to reconstruct the natural embedding of data onto a manifold and estimate the corresponding geodesic distance. Some of these are, for instance, the Isometric featuring mapping −Isomap− (Tenenbaum et al. [30]), Local Linear Embedding −LLE− (Roweis and Saul [27]), Laplacian Eigenmap (Belkin and Niyogi [3]), Hessian Eigenmap (Donoho and Grimes [10]), among others. In the following, we propose a robust version of the Isomap algorithm dedicated 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 temple 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 $t_{ij} \in \mathcal{T}$, with $\mathcal{T}$ a bounded interval of $\mathbb{R}$. For simplicity, we assume that all curves are observed at the same time with the same occurrence, i.e. $t_{ij} = t_j$ and $j = 1, \ldots, m$. Set $X_i = \{f_i(t_{ij}), j = 1, \ldots, m\} \in \mathbb{R}^m$ 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 $\mathcal{M} \subset \mathbb{R}^m$ (do Carmo [9]).

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 [23] . 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 $\alpha = 2$, if there exists, the minimizers of $F_\alpha(\mu)$ are called the Fréchet (or intrinsic) median and mean respectively. Following Koenker [19], 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.

The idea is to build a simple metric graph constructed with 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. [30], Bernstein et al. [4], and de Silva and Tenenbaum [8].

The construction of the weighted neighborhood graph in the first step of the Isomap algorithm and 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 versions does not exclude any point and does not require any additional tuning parameter free. 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. [12]. 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 $\mathcal{G}_{\mathrm{E}}$, $\sum_{\{X_i, X_{i'}\} \in E_{\mathrm{T}}} \mathrm{d}(X_i, X_{i'})$. The underlying idea in this construction is that, if two points X_i and $X_{i'}$ are relatively close, then we have that $\delta(X_i, X_{i'}) \approx \mathrm{d}(X_i, X_{i'})$. 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. [30], and de Silva and Tenenbaum [8]).

An approximation of $\delta(X_i, X_{i'})$ 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. $\hat{\delta}(X_i, X_{i'}) = \min_{g_{ii'} \in \mathcal{G}_{\mathrm{MST}}} L(g_{ii'})$, where $L(g_{ii'})$ 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 underying 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 $B(X_i, \epsilon_i) \subset \mathbb{R}^m$ around the point X_i with radius $\epsilon_i = \max_{\{X_i, X_i\} \in E_{\mathrm{T}}} \mathrm{d}(X_i, X_{i'})$. 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

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

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 $\overline{X_i X_{i'}} = \{X \in \mathbb{R}^m,\ \exists \lambda \in [0,1],\ X = \lambda X_i + (1-\lambda) X_{i'}\}$. Compute the shortest path distances between all pairs of points in the $\mathcal{G}'$ using Floyd's or Dijkstra's algorithm (see, e.g. Lee and Verleysen [20]).

4: Estimate the geodesic distance between two points by the length of the shortest path in the graph between these points.

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 [7].

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:

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

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 [25]. This is done in Section 6 for our real data set.

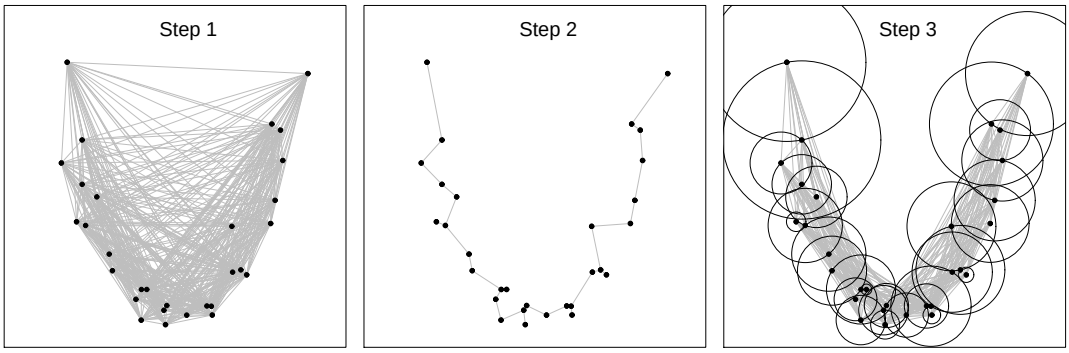

Figure 1: The 3-step construction of a subgraph $\mathcal{G}'$ from Simulation (4). On the left, the simulated data set (black dots) and the associated complete Euclidean graph $\mathcal{G}_{\mathrm{E}}$ (Step 1). On the middle, the $\mathcal{G}_{\mathrm{MST}}$ associated with the complete graph $\mathcal{G}_{\mathrm{E}}$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

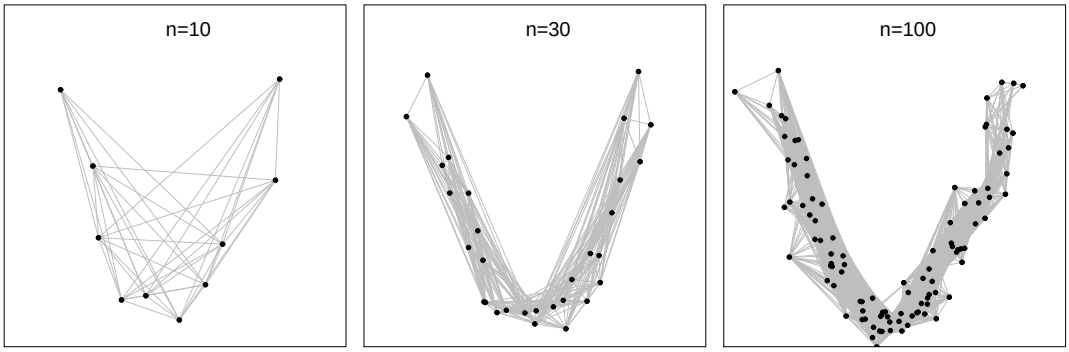

Figure 2: Evolution of graph $\mathcal{G}'$ from Simulation (4) for $n = 10, 30, 100$.

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 [22] and Dupuy et al. [11] propose to use as a good pattern of the dataset the so-called *structural expectation* 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. They build a registration procedure in order to estimate f_{SM}.

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 [9]). 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

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

The following theorem gives an minimizer.

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

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

where $\widehat{\text{med}}(A)$ 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. [12].

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 [21]). We also compare the estimators with the cross-sectional

median. The behavior of the estimator when the number of curves increases 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 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.

The Figure 3, illustrates the simulated data sets from Simulations 1-4 with $n = 30$ curves. For all of simulations, is clear that the cross-sectional median underperforms all other methods. For simulation one, where there is only phase variability, all of methods, except the median, 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 two. For simulations three and four, where there is additional amplitude variability, the best graphical results are achieved by the Isomap and robust manifold estimators. Note that in the simulation four, both of approaches coincide. On the contrary, the MBM estimator has an opposite pattern with respect to the theoretical template.

In order to compare more accurately the estimators described above, we calculate, for each one, the empirical mean squared error obtained on a set of $R = 30$ 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:

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

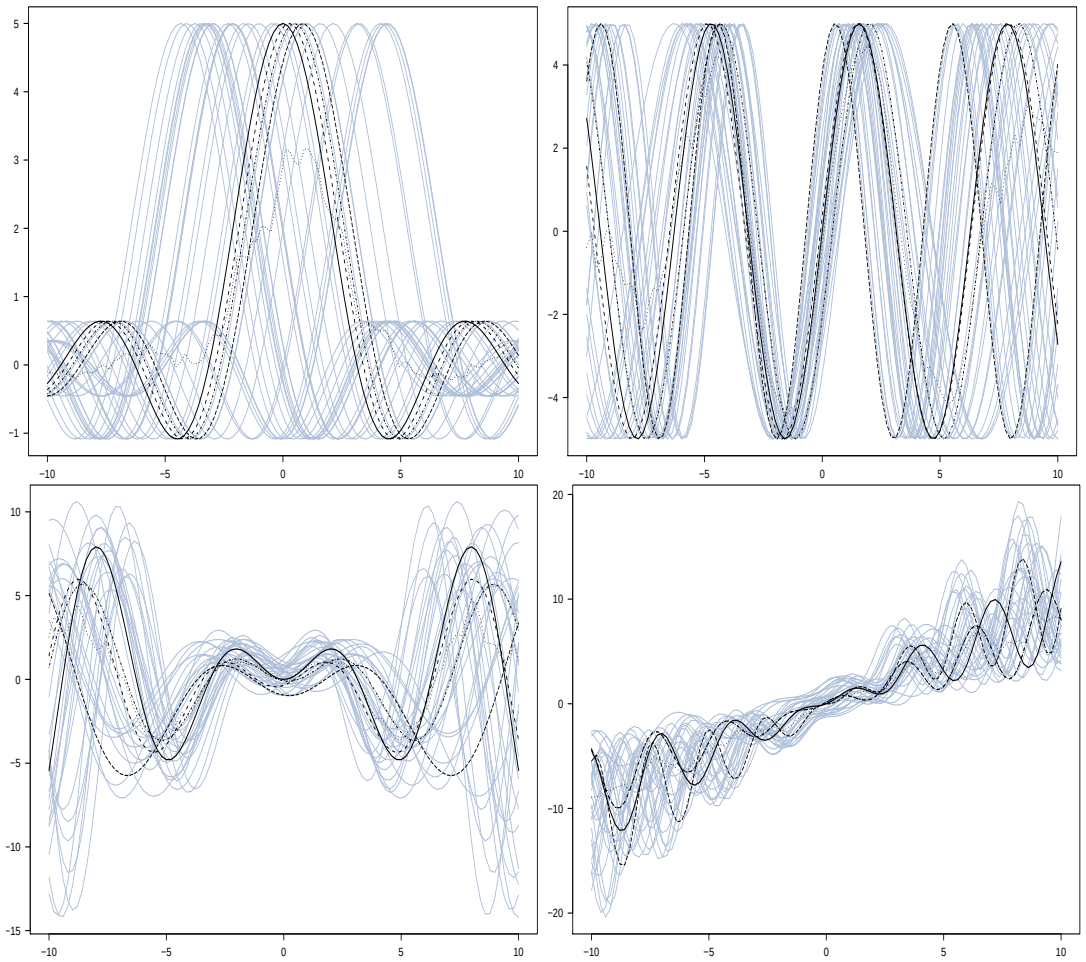

Figure 3: Simulated curves (gray) from simulation 1 (top left), 2 (top right), 3 (bottom left) and 4 (bottom right), the target template function f (black solid line), and cross-sectional median (dotted line), MBM (long dashed line), Isomap (dash-dotted line), and RME (dashed-bold line) template estimators.

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 $n = 15, 30, 45, 60$. 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 see that the cross-sectional median obtains some good results with respect to the other three methods, specially for the simulation 4. However, this apparent good behavior is not validated by the graphical results in figure 3, due to it does not approximate adequately the theoretical template. The template estimator based on the modified band median method has the worse results. As we expected, the values for the statistics calculated are bigger for more complex shape functions, it is when there are both amplitude and phase variations in the samples of curves.

Comparing only the results for the graph-based estimators, i.e. Isomap and RME, the smallest mean squared errors and variances are achieved by the template estimator calculated with our robust algorithm for all types of simulation considered and for all of sample size n, except in four and three cases, respectively (see Table 1). Also, it is almost true for the squared biases. 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 with different sample sizes.

n	Statistic	Simulation1				Simulation 2			
		Median	MBM	Isomap	EMS	Median	MBM	Isomap	EMS
15	MSE	**218**	416	236	223	341	966	**274**	398
	Bias2	167	180	63	**57**	169	229	**25**	56
	Variance	**51**	237	173	166	**172**	736	248	342
30	MSE	217	261	**83**	85	245	1234	274	**192**
	Bias2	189	109	18	**15**	157	378	38	**10**
	Variance	**28**	151	65	70	**88**	856	236	182
45	MSE	189	434	53	**50**	228	1035	247	**177**
	Bias2	171	183	**5**	16	167	268	16	**6**
	Variance	**18**	252	48	34	**61**	767	231	171
60	MSE	188	471	71	**34**	230	1413	200	**148**
	Bias2	176	227	6	**1**	177	543	14	**13**
	Variance	**13**	244	65	33	**53**	870	185	134
n	Statistic	Simulation 3				Simulation 4			
		Median	MBM	Isomap	EMS	Median	MBM	Isomap	EMS
15	MSE	**822**	2140	1163	994	**548**	1148	889	952
	Bias2	487	604	373	**274**	**466**	536	499	587
	Variance	**335**	1536	789	720	**82**	612	390	365
30	MSE	**705**	2812	762	836	**482**	1194	812	803
	Bias2	552	1207	**212**	340	**440**	498	477	484
	Variance	**153**	1605	550	496	**42**	696	335	319
45	MSE	607	2813	704	**488**	**494**	1103	893	878
	Bias2	514	1165	178	**174**	465	**445**	618	613
	Variance	**93**	1648	525	313	**29**	658	275	265
60	MSE	549	2015	456	**368**	**482**	960	786	730
	Bias2	472	444	**71**	119	**460**	504	481	447
	Variance	**77**	1571	386	249	**23**	456	305	283

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 [25]), 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 [25] for details).

Figure 4 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). From the curves 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.

In Figure 4, we see that for the case of sample acceleration curves of boys (bottom-right graph), the template of all methods coincide, selecting the same template curve which capture appropriately the common shape pattern. In the case of acceleration curves of girls (bottom-left graph), the three methods choose different functions. Nevertheless, the Isomap and Robust Manifold estimators have a similar behavior and MBM estimator is quite different. Note also that, although the conceptual idea of the MBM estimator is to search for a central function in the sample, it seems to choice curves that departs from the center in some time intervals, like in the velocity (in both girls and boys cases) and acceleration (in the case of girls) curves. For the velocity curves of girls, the Isomap and MBM methods choose the same curve, and in the case of boys our robust estimator and the Isomap procedure coincide. In general, we see that the RME seems to play a good work extracting a meaningful shape curve coinciding with at least one of the two other methods.

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 childrens through a gait cycle, where time is measured in terms of the child's gait cycle such that every curve is given for

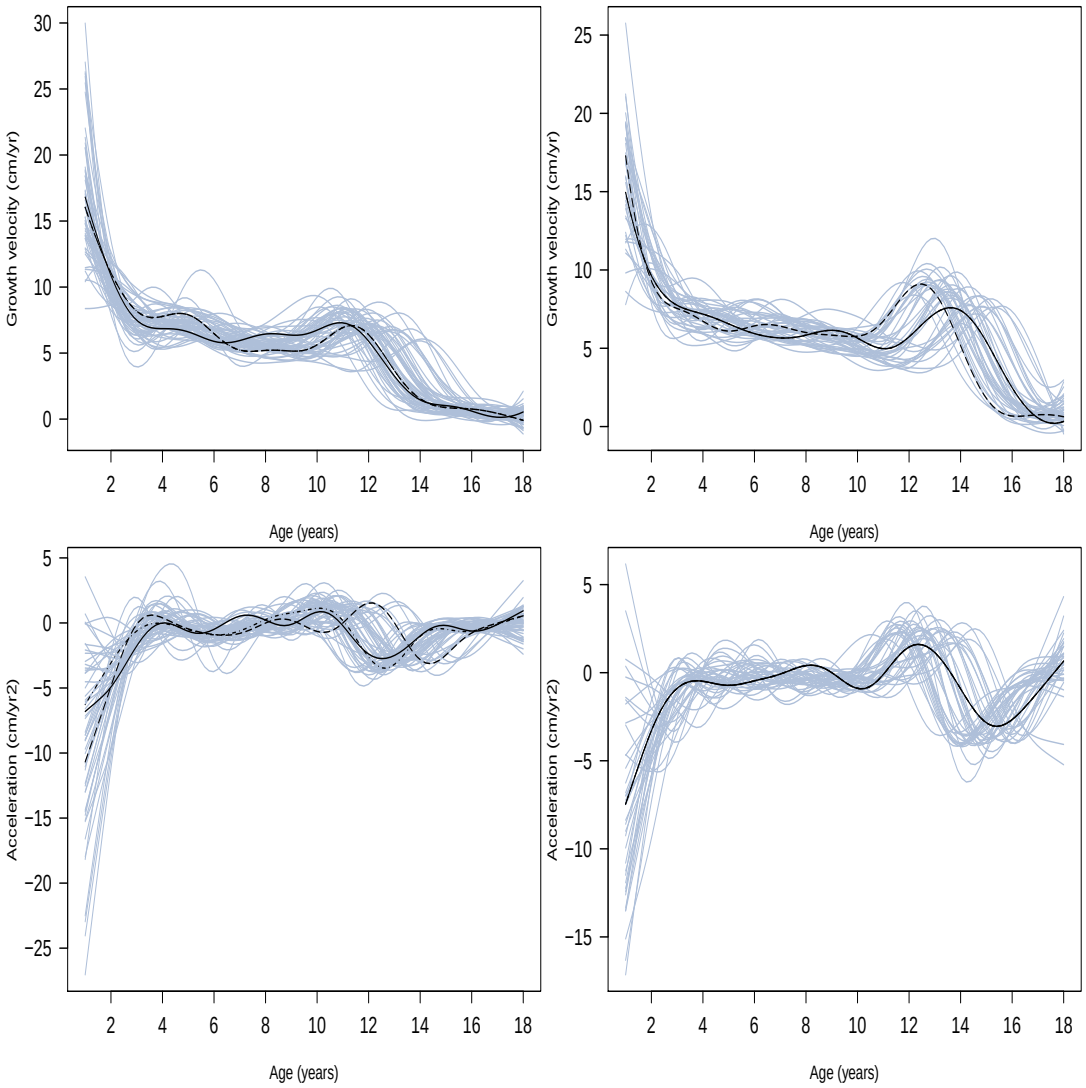

Figure 4: Velocity (on the top) and acceleration curves (on the bottom) of 54 girls (on the left) and 31 boys (on the right) in the Berkeley growth study (grey solid lines). The estimated template functions with MBM (long dashed line), Isomap (dashed-dotted line), and RME (solid line) methods.

values ranging between 0 and 1. The smoothed curves were obtained by fitting a Fourier basis system following the analysis of Ramsay and Silverman [25] for this data, where both sets of curves are periodic. Figure 5 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 curves. Note also that the same templates were chosen by the Isomap method. Although the MBM estimator choose a different curve as shape target function, it choose also a quite good representative for the hip and knee angle curves.

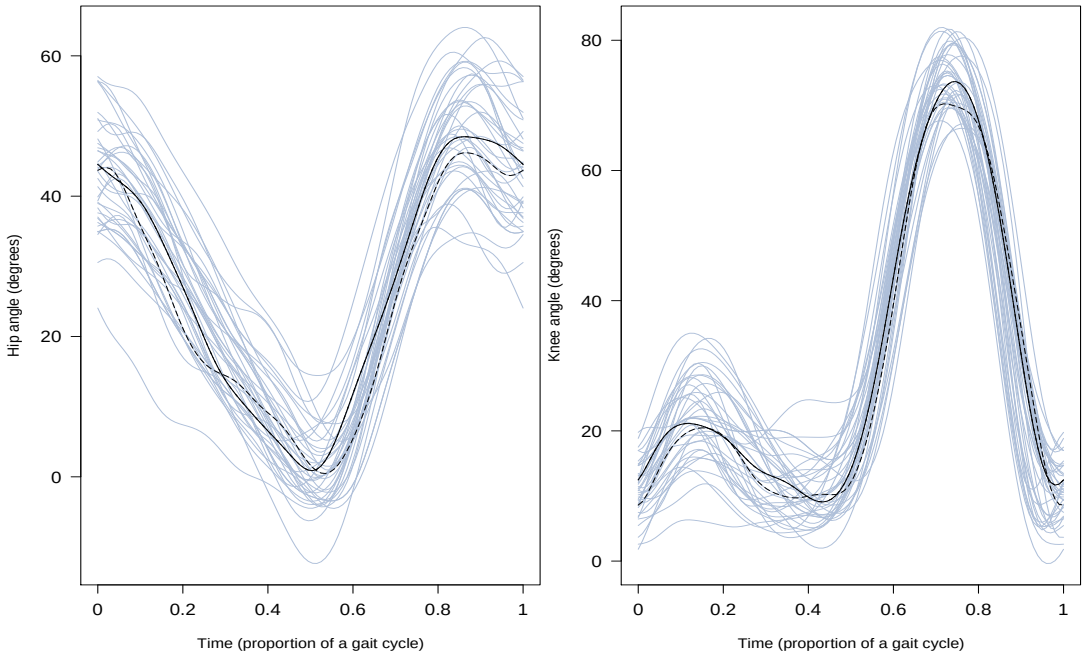

Figure 5: Angle curves formed by the hip (on the left) and knee (on the right) as 39 children go through a gait cycle

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 6, the smoothed curves corresponding to reflectance patterns of two 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 6, we observe that all of three estimators choose a different curve as representative function for both landscapes. In 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. The same

conclusions can be drawn for the wheat landscape, where the local extrema are well represented. 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.

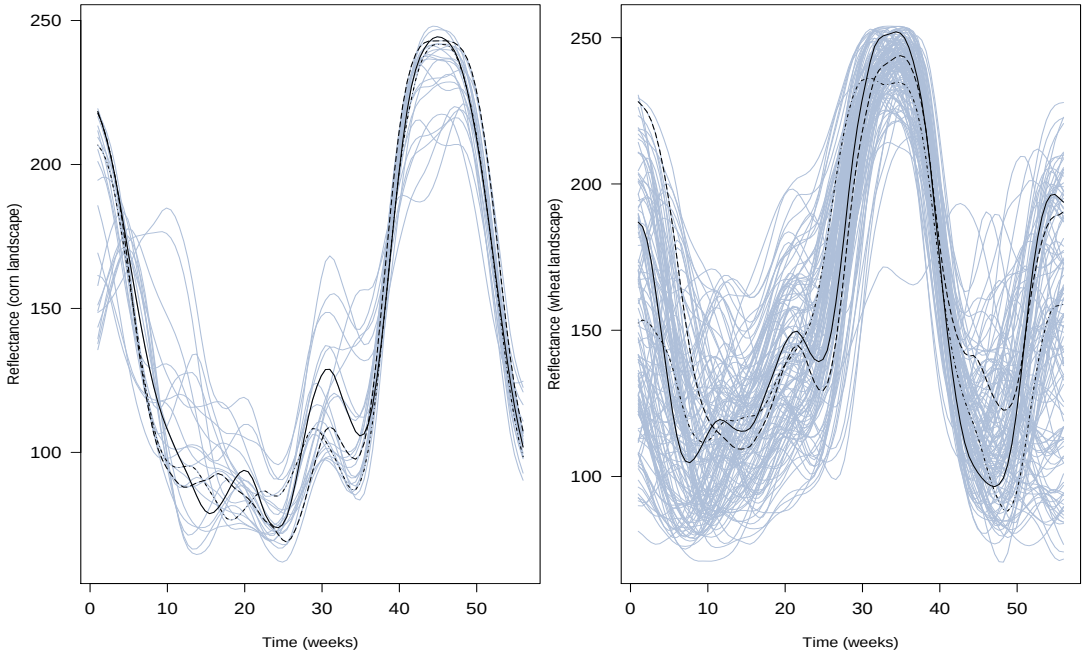

Figure 6: Reflectance curves of corn (left) and wheat (right) landscapes.

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 an empirical Fréchet median of the functions. This function is a meaningful template curve for a sample of functions, which has 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, we show that our method improves the performance of other pattern extraction methods, for simulated and real data sets.

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 the applications with some similar results with to respect to our algorithm. 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. [30]). This drawback is clearly related with the high variance of our estimator discussed previously and should be outperformed with further work. But, anyway, we show that our method improves the performance of other classical ones. The `R` code is available at the webpage of the authors 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 $\widehat{\text{med}}(A) = A_{(q+1)}$. Moreover, let $\alpha = A_{(j)}$ with $j < q + 1$. By symmetry, the case $j > q + 1$ holds. Then, we rewrite $C(\alpha)$ as

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

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

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

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

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

■

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 Computation*, 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. *Probability Theory and Related Fields*, 152:425–473, 2012.

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

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

[8] 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.

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

[10] D. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100:5591–5596, 2003.

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

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

[13] F. Gamboa, J. Loubes, and E. Maza. Semi-parametric estimation of shits. *Electronic Journal of Statistics*, 1:616–640, 2007.

[14] G. James. Curve alignment by moments. *The Annals of Aplied Statistics*, 1: 480–501, 2007.

[15] A. Kneip and T. Gasser. Convergence and consistency results for self-modelling regression. *The Annals of Statistics*, 16:82–112, 1988.

[16] A. Kneip and T. Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20:1266–1305, 1992.

[17] A. Kneip and J. Ramsay. Combining registration and fitting for funstional models. *Journal of the American Statistical Association*, 103:1155–1165, 2008.

[18] A. Kneip, X. Li, X. MacGibbon, and J. Ramsay. Curve registration by local regression. *Canadian Journal of Statistics*, 28:19–29, 2000.

[19] R. Koenker. The median is the message: Toward the Fréchet median. *Journal de la Société Française de Statistiques*, 147:61–64, 2006.

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

[21] S. López-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104:718–734, 2009.

[22] E. Maza. Estimation de l'espérance structurelle d'une fonction aléatoire. *Comptes Rendus Mathématique. Académie des Sciences. Paris*, 343, 2006.

[23] X. Pennec. Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

[24] J. O. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society. Series B*, 60:351–363, 1998.

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

[26] B. Rønn. Nonparametric maximum likelihood estimation for shifted curves. *Journal of the Royal Statistical Society. Series B*, 63:243–259, 2001.

[27] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[28] B. W. Silverman. Incorporating parametric effects into functional principal components analysis. *Journal of Royal Statistical Society. Series B*, 57:673–689, 1995.

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

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

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

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

[33] K. Wang and T. Gasser. Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25:1251–1276, 1997.