The Hidden Manifold Distance for Functional Data

Abstract

In many machine learning applications, the observed data is a discretely-sampled realisation of a continuous process. While one can naively treat such data as vectors in Euclidean space, one may do better to regard them as functional data. When the underlying curves of interest lie on a manifold, e.g. probability density functions or warped curves of a common template function, there is further structure to exploit. This work addresses the estimation of pairwise geodesic distances between functional manifold data that are observed with noise, causing them to lie off the manifold. This setting falls outside of classic manifold learning techniques which require data to live exactly on or very near a manifold. The proposed methodology first sends the observed functional data to the hidden manifold, estimated using subspace-constrained mean-shift. Geodesic distances are subsequently calculated by employing shortest-path algorithms on this estimated manifold. Improved estimation of the pairwise geodesic distance has potential benefits for downstream tasks such as distance-based functional classification.

Introduction

Many statistical and machine learning methods rely on some measure of distance. For example, clustering and classification of functional data, discretely-sampled curves varying over a continuum, is a common learning task in which distance plays a crucial role. In a naive treatment, the data may be simply processed as vectors in Euclidean space. Doing so however risks losing the richness of functional data; indeed, one may do better to regard the data as functions. The challenge of analyzing infinite-dimensional objects can be ameliorated if the curves are smooth since the dimensionality is then only artificially high. This insight underpins functional data analysis, a subfield of statistics that studies functional data.

In this work, we advocate that for functional manifold data, i.e. functional data that lie on a manifold, the geodesic distance is more appropriate than the L^2 distance. Indeed, when the manifold hypothesis is plausible, e.g. for classes of probability density functions and classes of warped curves of a common template function, clustering and classification using the geodesic distance may give better results than using the L^2 distance.

The manifold setting presents certain challenges since, for nonlinear manifolds, even basic operations such as addition and subtraction require special consideration. We will carefully lay out a technique for estimating the geodesic distance between noisy functional manifold data.

Let $X_1, \ldots, X_n$ be a sample of n independent realizations of a random variable X that takes value in the Hilbert space $L^2([a,b],\mathbb{R})$. Suppose additionally that the function X belongs to a Riemannian manifold $\mathcal{M} \subset L^2([a,b],\mathbb{R})$ with Riemannian metric tensor g which are often used to assign a metric on the manifold as follows (Lin et al. 2014). For each point X on the manifold, the Riemannian metric tensor g has an inner product g_X on the tangent space $T_X\mathcal{M}$. The norm of a tangent vector $V \in T_X\mathcal{M}$ is defined as

$$\|V\| = \sqrt{g_X(V,V)}.$$

The geodesic distance between two functions X_i, X_j on the manifold $\mathcal{M}$, based on this metric tensor g, is defined as

$$d_{\mathcal{M}}(X_i, X_j) := \inf_{\gamma:[0,1]\to\mathcal{M},\gamma(0)=X_i,\gamma(1)=X_j} l(\gamma),$$

where

$$l(\gamma) := \int_0^1 \left\| \frac{d\gamma}{dt}(t) \right\| dt$$

is the length of the piecewise-smooth curve γ.

Ideally, the functions X_i and X_j are observed on a very dense domain grid with no measurement error. Then the geodesic distance $d_{\mathcal{M}}(X_i, X_j)$ can be estimated using any of a variety of shortest-path algorithms, e.g. the Floyd-Warhsall (Floyd 1962) algorithm. However, most shortest-path algorithms critically assume observations live near the manifold with little noise (Yeh et al. 2005) and are thus likely to fail for functional data which can be observed with low signal-to-noise ratio. In this work, we put forth a technique for the recovery of the $n \times n$ matrix G of pairwise geodesic distances:

$$G(i,j) = G(j,i) = \begin{cases} d_{\mathcal{M}}(X_i, X_j) & \text{if } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

In lieu of X_i and X_j, we have access only to discretely-sampled noisy functional observations Y_i and Y_j that possibly lie off the true manifold.

The work of (Chen and Muller 2012) was among the first in functional data analysis to consider the manifold hypothesis for functional data. A notion of the mean and variation of functional manifold data was introduced there. Of particular relevance to this work is their proposed P-ISOMAP procedure which allows for noisy functional observations. This is in contrast to the classic ISOMAP algorithm (Tenenbaum, Silva, and Langford 2000) which assumes observations lie exactly on the manifold. This drawback to ISOMAP was also realised in (Dimeglio et al. 2014) who proposed a procedure we will call robust-ISOMAP for functional data that is less sensitive to outliers. We will discuss P-ISOMAP and robust-ISOMAP in more details below.

Proposed method for estimating geodesic distances

Suppose that each curve $X_i \in (\mathcal{M}, g)$ is observed with measurement error on a time grid $T_i = (t_{i1}, \ldots, t_{iK}), a \leq t_{i1} < \ldots < t_{iK} \leq b$, i.e. we observe a sample of K-dimensional vectors $Y_1, \ldots, Y_n$ with $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$. Let the random variables ϵ_{ij} have mean zero and be uncorrelated with each other. We assume that each observation grid $T_1, \ldots, T_n$ is dense, i.e. K is large.

We begin by converting the discretely-sampled functional observations into continuous ones. Let $\tilde{X}_1, \ldots, \tilde{X}_n$ denote the functional versions of the raw data obtained by some smoothing method. For example, we may employ spline smoothing (Ramsay et al. 2005) to recover the functional versions of the raw data, i.e.

$$\tilde{X}_i = \arg \min_{f \in C^2[0,1]} \left\{ \sum_{j=1}^{K} \left(f(t_{ij}) - Y_{ij}\right)^2 + \lambda \|\partial_t^2 f\|_{L^2}^2 \right\}, \tag{1}$$

where $\lambda > 0$ is a tuning parameter controlling the smoothness of $\tilde{X}_i$. Our method is based on the idea that the underlying functional manifold $\mathcal{M}$ can be sufficiently well-recovered by the subspace-constrained mean-shift (SCMS) algorithm (Ozertem and Erdogmus 2011) which we now review briefly in the Euclidean case. Suppose $Z_1, \ldots, Z_n$ is a sample of realisations of a random vector $Z \in \mathcal{M}$. We can use a kernel density estimator $\hat{f}$ based on this data to estimate the probability density function of Z. Let the set $\hat{M}$ be comprised of the basins of attraction of $\hat{f}$. The SCMS algorithm sends each point Z_i to its destination in $\hat{M}$. Theoretical justification that $\hat{M}$ as estimated by SCMS is a reasonable surrogate for the true manifold $\mathcal{M}$ can be found in (Genovese et al. 2014).

Before presenting our procedure, we also need to review the ISOMAP algorithm (Tenenbaum, Silva, and Langford 2000), a three-step procedure that takes as input a set of points $x_1, \ldots, x_n$ in a submanifold $\mathcal{M} \subset \mathbb{R}^D$ and produces an embedding in the space $\mathbb{R}^d$ with $d < D$ that preserves pairwise geodesic distances. The procedure is as follows.

1. Construct a weighted graph $\mathcal{G}$ with nodes corresponding to the observations $x_1, \ldots, x_n \in \mathbb{R}^D$. Two nodes x_i and x_j are connected by an edge e_{ij} with weight $d_{ij}1\{d_{ij} \leq \varepsilon\}$ where $d_{ij} = \|x_i - x_j\|_2$ and $\varepsilon > 0$.
2. Estimate the pairwise geodesic distances $d_{\mathcal{M}}(x_i, x_j)$ based on $\mathcal{G}$ using shortest-path algorithms. Specifically, this is the sum of the weights of the edges forming the shortest path, which is calculated either with the Floyd-Warshall algorithm (Floyd 1962) or with the Dijkstra algorithm (Dijkstra 1959).
3. Use multidimensional scaling to obtain an embedding in $\mathbb{R}^d$ that preserves the pairwise geodesic distances estimated above.

Note that shortest-path algorithms such as the Floyd-Warshall algorithm are used to find the smallest path *given* a weighted graph but they do not produce the weighted graphs themselves. Both P-ISOMAP and robust-ISOMAP are modifications of ISOMAP in the sense that they change the construction of the weighted graph in step 1 above.

In what follows, we call IsoGeo the two-step procedure which performs a modified version of the first step of ISOMAP followed by the original second step of ISOMAP. The first step of IsoGeo constructs a weighted graph $\mathcal{G}$ according to (Dimeglio et al. 2014) as follows. First we construct a complete weighted graph $\mathcal{G}_c$, meaning that every pair of nodes x_i and x_j is connected by an edge of weight d_{ij}, so that the graph is independent of the tuning parameter ε. Next, we obtain the minimal spanning tree associated with $\mathcal{G}_c$ and denote its set of edges by E_s. The graph $\mathcal{G}$ is obtained by adding all edges e_{ij} to E_s for which the following condition is true:

$$\overline{x_i x_j} \subset \bigcup_{i=1}^{n} B(x_i, \varepsilon_i),$$

where $B(x_i, \varepsilon_i)$ is the open ball of center x_i and radius $\epsilon_i = \max_{e_{ij} \in E_s} d_{ij}$, and $\overline{x_i x_j} = \{x \in R^D \mid \exists \lambda \in [0,1], x = \lambda x_i + (1-\lambda) x_j\}$. Let the distance estimated by IsoGeo be denoted IsoGeo(x_i, x_j). Note that IsoGeo has no tuning parameters.

We are now ready to describe our procedure. Since SCMS is based on kernel density estimation, we first reduce the dimension of our data with multidimensional scaling before applying SCMS to avoid the curse of dimensionality.

1. Discretise the functions $\tilde{X}_1, \ldots, \tilde{X}_n$ by evaluating them on a dense common grid $\tilde{T} = (t_1, \ldots, t_D)$. Let s be a positive integer, much smaller than D. Obtain $\tilde{X}_1^s, \ldots, \tilde{X}_n^s \in \mathbb{R}^s$ using multidimensional scaling so that the pairwise $\mathbb{R}^D$ Euclidean distances of the discretised functions $\tilde{X}_i$ can be preserved as much as possible.
2. Apply the subspace constrained mean-shift algorithm (Ozertem and Erdogmus 2011) to each of $\tilde{X}_i^s$ to obtain its destination $\tilde{X}_i^{s,\hat{\mathcal{M}}}$.
3. Our geodesic distance estimator is the $n \times n$ matrix $\hat{G}$ whose elements are given by

$$\hat{G}(i,j) = \hat{G}(j,i) = \begin{cases} \text{IsoGeo}(\tilde{X}_i^{s,\hat{\mathcal{M}}}, \tilde{X}_j^{s,\hat{\mathcal{M}}}) & \text{if } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

There are two tuning parameters to our procedure. First, there is the dimension s in step 1. In our simulations we try $s \in \{1, 2, 3\}$. The other parameter is the bandwidth h in the subspace constrained mean-shift. In our simulations we set h according to the heuristic in Equation (A1) of (Chen et al. 2015):

$$h = \left(\frac{4}{s+2}\right)^{\frac{1}{s+4}} n^{\frac{-1}{s+4}} \sigma_{\min},$$

where $\sigma_{\min}$ is the minimum over $j = 1, \ldots, s$ of the standard deviation of $\tilde{X}^s_{1,j}, \ldots, \tilde{X}^s_{n,j}$, with $\tilde{X}^s_{i,j}$ being the jth entry of the vector $\tilde{X}^s_i$. In reality, depending on the downstream task, both s and h can be tuned in a data-adaptive way, say using cross-validation.

Simulation study

We perform simulations to study various estimators of pairwise geodesic distances. We compare the performance of our method to each of the following four alternatives. The first alternative, denoted **RD**, takes the naive approach of applying IsoGeo directly on the raw vectors $Y_1, \ldots, Y_n \in R^K$. Note that this procedure is only sensible if all the time grids $T_1, \ldots, T_n$ are the same.

For the next two alternatives, we first smooth the discretely-sampled noisy functional data using splines as described in (1). Then, we either estimate pairwise distance between the smoothed $\tilde{X}_1, \ldots, \tilde{X}_n$ evaluated on a common time grid of D points using IsoGeo or the L^2 distance. The former method will be denoted **Spline IsoGeo** and the latter **Spline L2**.

The final alternative we consider is P-ISOMAP which is the two-step procedure developed in (Chen and Muller 2012) for handling noisy functional data. In the first step, a penalty is incorporated to robustify the construction of the weighted graph. This is followed by the original second step of ISOMAP.

The code for P-ISOMAP was provided by (Chen and Muller 2012) in MATLAB. We ran P-ISOMAP and tuned for the parameter ε and the penalty parameter δ using the relative Frobenius assessment metric, described below. We found that the penalty parameter chosen in this oracle fashion was always zero, thus reducing the P-ISOMAP procedure to standard ISOMAP. Given this, we did not make further comparison to P-ISOMAP in our simulation study. We also do not present the results for the RD method which always produced poor estimates of the geodesic distance.

Three different metrics are used to assess the quality of a pairwise geodesic distance estimator. The first metric assesses near-ϵ isometry, which we define as the percentage of estimated pairwise geodesic distances between $1 - \epsilon$ and $1 + \epsilon$ of the corresponding true pairwise geodesic distance. We form the receiver operating curve with ϵ on the x-axis and near-ϵ isometry on the y-axis. The metric is the area under this receiver operating curve. Thus a good estimator has high values for the near-isometry metric.

The second metric $\|G - \hat{G}\|/\|G\|$ is the Frobenius norm of the estimation error matrix $G - \hat{G}$, relative to the Frobenius norm of the true distance matrix G. The third metric is the Pearson correlation coefficient between the upper diagonal

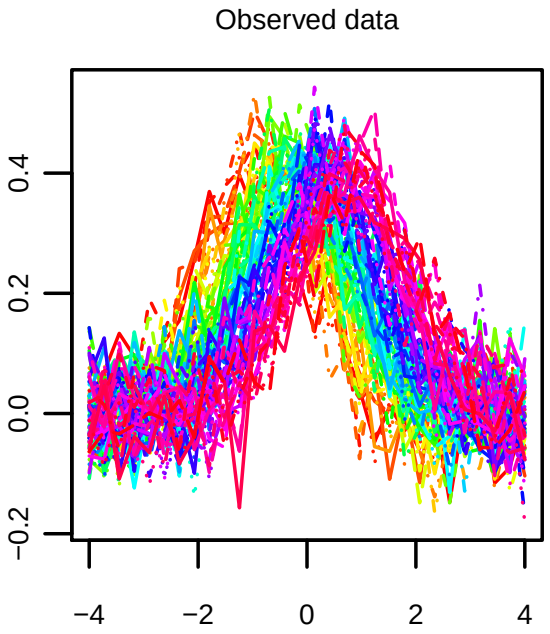

Figure 1: One hundred noisy curves in the functional manifold of normal probability density.

of G and $\hat{G}$. A good geodesic distance estimator has low Frobenius norm and high Pearson correlation coefficient.

We shall consider three functional manifolds in our simulation study. Throughout, we make use of a helpful property of one-dimensional manifolds. If $X \in \mathcal{M}$, with the L^2 metric tensor, depends on a scalar parameter $\theta \in \mathbb{R}$, then

$$d_{\mathcal{M}}(X_{\theta_1}, X_{\theta_2}) = \int_{\theta_1}^{\theta_2} \left\| \frac{\partial X}{\partial \theta}(t) \right\|_{L^2} d\theta.$$

We also need to make a disclaimer about our sampling technique below. Proper sampling from a manifold remains an open problem. Even in the Euclidean case, it is not obvious how sampling should be done (Diaconis, Holmes, and Shahshahani 2013). For now, to safeguard against sampling unevenly on the functional manifold, we sample on a very narrow support of the manifold parameters.

Isometric functional manifold of normal density functions

We take Manifold 2 in (Chen and Muller 2012) and set the variance of the normal density to be 1. Let $X_\beta : [a, b] \to \mathbb{R}$ be given by $X_\beta(t) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(t - \beta)^2\right]$. Consider

$$\mathcal{M} = \{X_\beta : \beta \in [-1, 1], t \in [a, b]\},$$

with the L^2 inner product as the metric tensor. First, note that the "straight" line connecting X_{β_1} and X_{β_2} does not always stay inside of $\mathcal{M}$. Thus the geodesic distance will not coincide with the L^2 distance.

We set $a = -4$ and $b = 4$ and sample β according to a truncated standard normal with support on $[-1, 1]$. The

geodesic distance between the curves X_{β_1} and X_{β_2} is given by

$$\begin{aligned} d_{\mathcal{M}}(X_{\beta_1}, X_{\beta_2}) &= \int_{\beta_1}^{\beta_2} \left\| \frac{dX_\beta(t)}{d\beta} \right\|_{L^2} d\beta \\ &= \int_{\beta_1}^{\beta_2} \sqrt{\frac{1}{2\sqrt{\pi}} \int_{-4}^{4} (t-\beta)^2 f(t) dt}\, d\beta \\ &\approx \int_{\beta_1}^{\beta_2} \sqrt{\frac{1}{2\sqrt{\pi}} \frac{1}{2}}\, d\beta \\ &= (\beta_2 - \beta_1) \frac{1}{2\pi^{1/4}}, \end{aligned}$$

where f is the density of a $\mathrm{N}(\beta, 1/2)$ and the approximation in the next to last step comes from integrating on $[a, b] = [-4, 4]$ rather than $\mathbb{R}$. We can see this manifold is isometric, since the geodesic distance between X_{β_1} and X_{β_2} in $\mathcal{M}$ is proportional to $\beta_2 - \beta_1$.

Functional manifold of square root velocity functions

It was shown in (Joshi, Srivastava, and Jermyn 2007) that the square root representation of probability density functions gives rise to a simple closed-form geodesic distance. They consider the manifold

$$\mathcal{M} = \{\psi : [0,1] \to \mathbb{R} : \psi \geq 0, \int_0^1 \psi^2(s)\, ds = 1\},$$

with the Fisher-Rao metric tensor

$$< v_1, v_2 >= \int_0^1 v_1(s) v_2(s)\, ds,$$

for two tangent vectors $v_1, v_2 \in T_\psi(\mathcal{M})$. Note that this coincides with the $L^2[0,1]$ inner product. It was shown in (Joshi, Srivastava, and Jermyn 2007) that the geodesic distance between any two ψ_1 and ψ_2 in $\mathcal{M}$ is simply

$$d_{\mathcal{M}}(\psi_1, \psi_2) = \cos^{-1} < \psi_1, \psi_2 > .$$

We will specifically examine the square root of $Beta(\alpha, \beta)$ densities, which are of course supported on $[0, 1]$. That is, consider

$$\mathcal{M} = \left\{ \sqrt{\psi_{\alpha,\beta}} : 1 \leq \alpha \leq 5, 2 \leq \beta \leq 5 \right\},$$

where $\psi_{\alpha,\beta} : [0,1] \to \mathbb{R}$ is the probability density function of a Beta random variable with shape parameters α and β. We sample α according to a truncated normal with mean 3 and variance 0.09 with support on $[1, 5]$. We sample β according to a truncated normal with mean 3.5 and variance 0.09 with support on $[2, 5]$.

Functional manifold of warped functions

Let $X_\alpha(t) = \mu(h_\alpha(t))$ be defined on $[-3, 3]$ where

$$\mu(t) = \exp\{(t - 1.5)^2/2\} + \exp\{(t + 1.5)^2/2\},$$

and

$$h_\alpha(t) = 6 \frac{\exp\{\alpha(t+3)/6\} - 1}{\exp\{\alpha\} - 1},$$

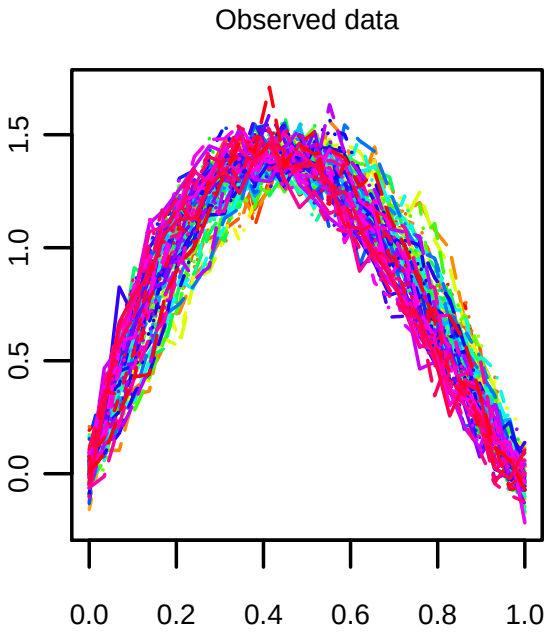

Figure 2: One hundred noisy curves in the functional manifold of square root probability density functions.

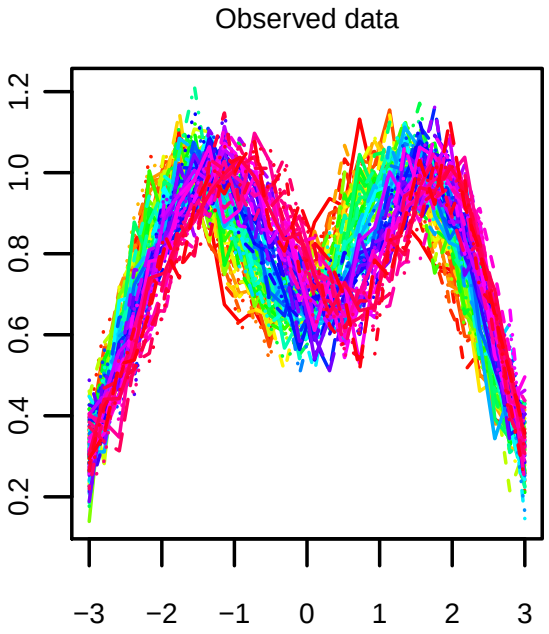

Figure 3: One hundred noisy curves in the functional manifold of warped functions.

if $\alpha \neq 0$ and $h_\alpha(t) = t$ otherwise.

Consider the manifold

$$\mathcal{M} = \{X_\alpha : -1 \leq \alpha \leq 1\}.$$

We sample α according to a truncated standard normal with support on $[-1, 1]$. This manifold is based on Equations 17 and 18 in (Kneip and Ramsay 2008) but with z_{i1}, z_{i2} therein both set to 1. (Note that Equation 17 therein has a typo where the exponentials are missing negative signs.)

The geodesic distance is

$$d_{\mathcal{M}}(X_{\alpha_1}, X_{\alpha_2}) = \int_{\alpha_1}^{\alpha_2} \left\| \frac{dX_\alpha(t)}{d\alpha} \right\|_{L^2} d\alpha.$$

We estimate this using numerical integration.

Results

For each manifold $\mathcal{M}$ described above, we simulate 100 Monte Carlo samples of 100 functions $X_i \in \mathcal{M}$. We then construct vectors $Y_i \in \mathbb{R}^K$ such that $Y_{ij} = X_i(t_j) + \epsilon_{ij}$ with $T = (t_1, \ldots, t_K)$ a common grid of equally spaced points and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The tuning parameters for the proposed method, s and h, were chosen as described at the end of the methodology section. In step 1 of our procedure, we take $D = 100$ for the size of the common time grid. We will investigate how two parameters affect the performance of the various methods in consideration. The first one is the signal-to-noise ratio

$$SNR = 10log_{10}\left(\frac{\sigma_{sig}^2}{\sigma_\epsilon^2}\right),$$

where σ_{sig}^2 is the variance of the signal. We fix SNR at either 0.1 (low) or 0.5 (high). The second parameter is K, the dimension of the discretely-observed noisy Y_i's. We set K to either 30 (sparse) or 100 (abundant).

We will present the results for the high signal-to-noise and sparse K setting. We show the performance of the Spline IsoGeo and Spline L2 method versus our method for three dimensions $s = 1, 2, 3$. We can see from Figures 4 and 5 that spline smoothing followed by either IsoGeo or the L^2 distance estimates the pairwise geodesic distances reasonably well though not as well as the proposed method for $s = 2$ and $s = 3$. Notably, when s in our method is not selected properly, for instance when $s = 1$, our estimator performs poorly in terms of the relative Frobenius norm although still better than Spline IsoGeo and Spline L2 in the other two metrics. It may be possible that for certain downstream tasks such as clustering or classification, precise estimation of the geodesic distance is less important than one that respects the ordering of the distances. In such cases, the near-isometry and Pearson correlation metrics may give a better indication of downstream performance of the geodesic distance estimator.

The result of the square root velocity manifold scenario is presented in Figure 6. The Spline L2 method has similar performance to our proposed method. This could be due to the fact that we did not sample the manifold broadly enough so that the sample lives in a locally flat region where the L^2 distance holds.

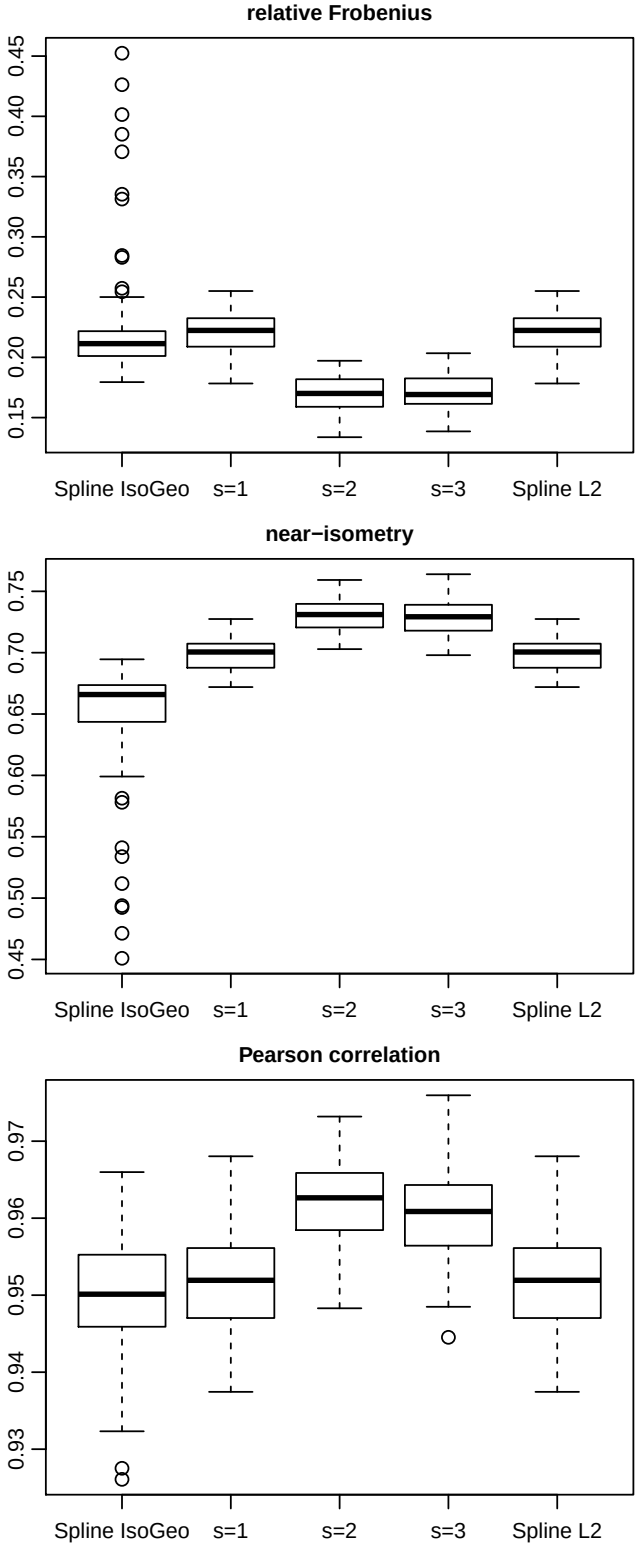

Figure 4: Manifold of $N(\beta, 1)$ probability density functions. The boxplots indicate the distributions of each of the three metrics over 100 Monte Carlo samples. Across all three metrics, our method with $s = 2$ performs the best, followed by our method with $s = 3$.

Distance-based functional classification

There are many downstream analysis tasks in functional data analysis that are based on pairwise distances, including

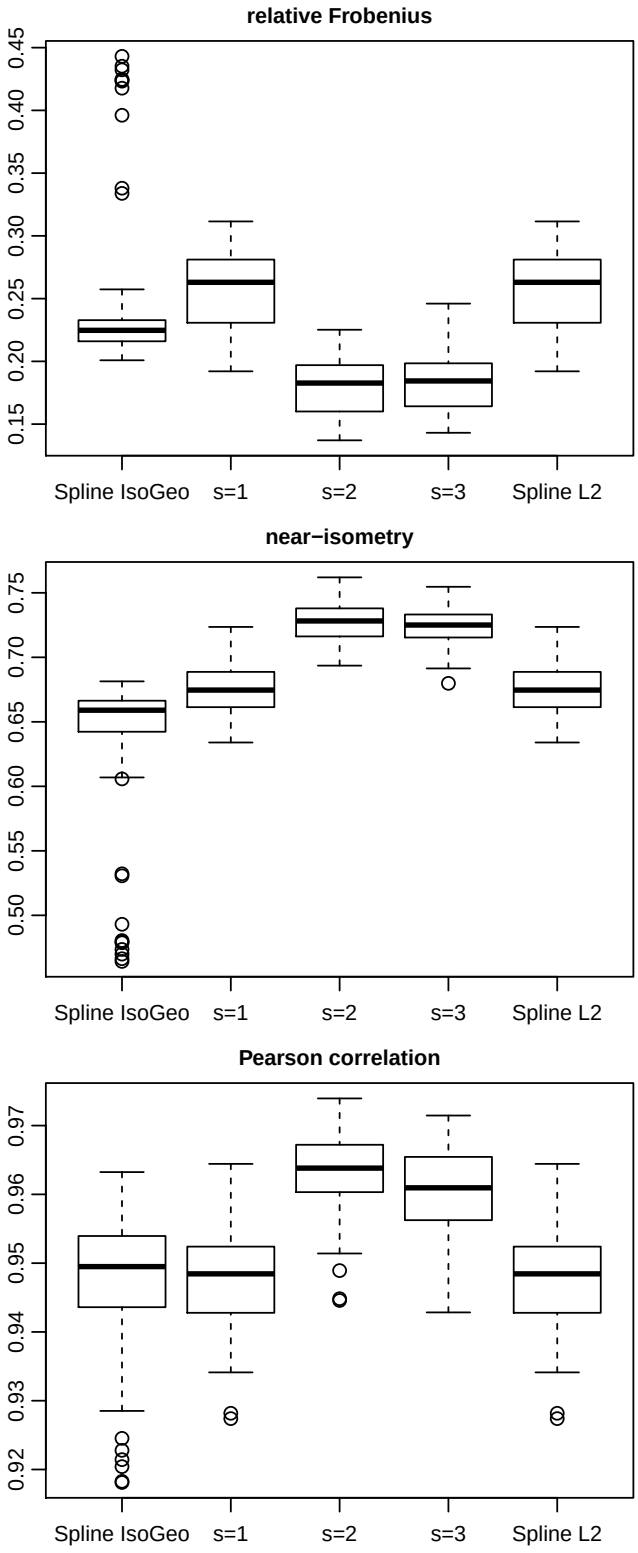

Figure 5: Manifold of warped functions. The boxplots indicate the distributions of each of the three metrics over 100 Monte Carlo samples. The best performing estimators are given by our method with $s = 2$ and then $s = 3$.

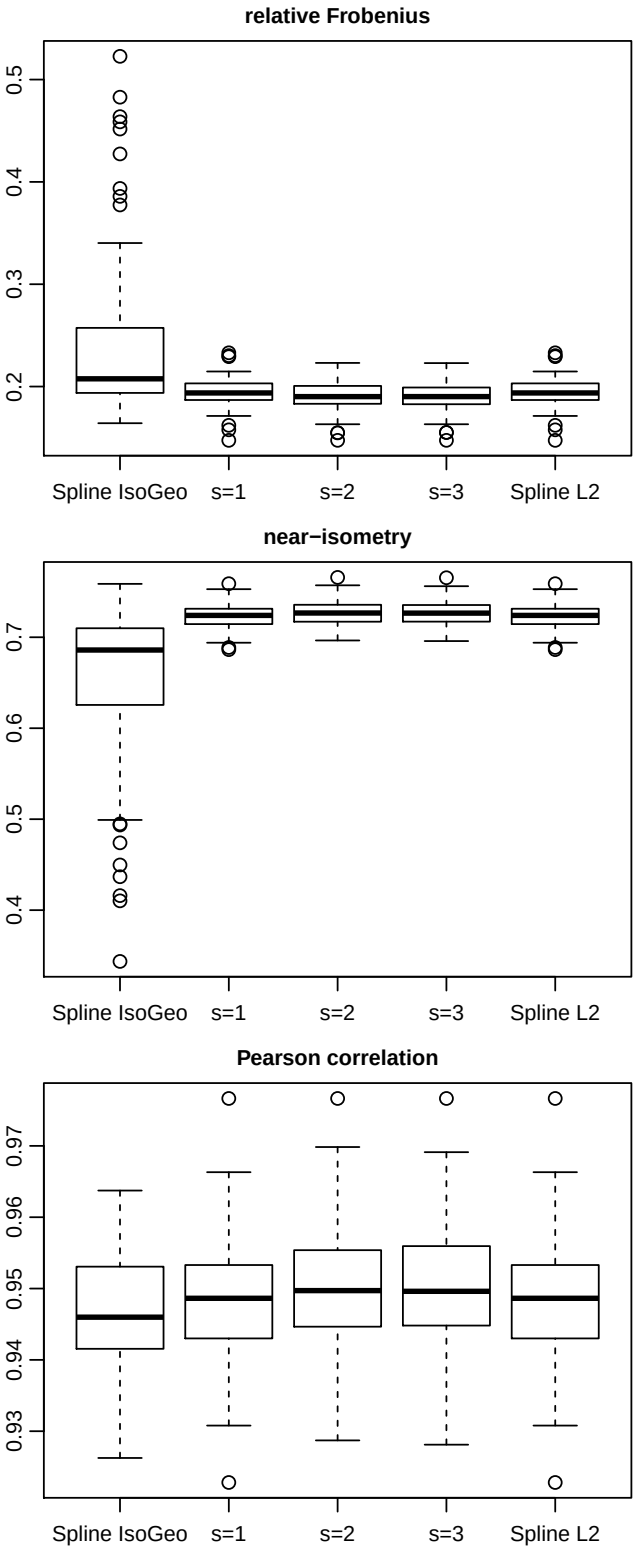

Figure 6: Manifold of square root Beta densities. The boxplots indicate the distributions of each of the three metrics over 100 Monte Carlo samples. Besides Spline IsoGeo, all methods perform comparably well.

distance-based nonparametric regression and distanced-based functional clustering. In this section, we explore whether our geodesic distance estimator has benefits for the downstream analysis task of functional classification.

We work with the non-parametric classifier proposed in (Ferraty and Vieu 2006) which is a functional version of the Nadaraya-Watson kernel estimator. Given a sample of curves $X_1, \ldots, X_n$ where each function X_i is associated to a class label Y_i, the estimator of the probability that a new observation $X^\star$ is in class $Y^\star = y$ is defined as

$$\hat{P}(Y^\star = y | X^\star) = \frac{\sum_{i=1}^n k[h^{-1} d(X^\star, X_i)] \mathbf{1}(Y_i = y)}{\sum_{i=1}^n k[h^{-1} d(X^\star, X_i)]}, \tag{2}$$

where d is a distance, k a kernel, h a bandwidth and $\mathbf{1}$ the indicator function. The curve $X^\star$ is then classified in the class $Y^\star = y$ for which the conditional probability (2) is largest. We compare the classification performance obtained with the distance d given by our proposed geodesic distance with $s = 1, 2$ and 3, as well as that given by Spline IsoGeo and Spline L2 described in the simulation study section. We set k to be a quadratic kernel and the bandwidth h is chosen by cross-validation.

The classification task is performed on the well-known Berkeley growth curves dataset in functional data analysis, available in the *fda* package (Ramsay, Hooker, and Graves 2009) for R (R Core Team 2017). We chose this dataset because we expect the data to contain nonlinear features as suggested in (Chen and Muller 2012). The data consist of the height of $n = 93$ individuals, 39 boys and 54 girls, measured on a common grid $t_1, \ldots, t_K$ of $K = 31$ points taken between the ages of 1 and 18 years. The raw data $Y_1, \ldots, Y_n$ are transformed to continuous functions by smoothing with B-spline bases with a roughness penalty on the fourth derivative :

$$\tilde{X}_i(t) = \sum_{l=1}^{b} c_{il} B_l(t),$$

where

$$(c_{i1}, \ldots, c_{ib}) = \arg\min_{c \in \mathbb{R}^b} \left\{ \sum_{j=1}^{K} \left(\sum_{l=1}^{b} c_{il} B_l(t_j) - Y_{ij} \right)^2 + \lambda \left\| \sum_{l=1}^{b} c_{il} \partial_t^4 B_l \right\|_{L^2}^2 \right\}.$$

We use $b = 35$ B-splines bases of order 6 and the parameter λ is chosen by generalized cross-validation. We then differentiate the resulting fonctions to obtain velocity curves; the data are illustrated for boys and girls on Figure 7.

To assess the classification performance of the the proposed geodesic distance, we randomly split 200 times the dataset into a training set of 50 curves and a test set of 43 curves. For each split, we calculate the misclassification error, and we average these errors over the splits. The results are presented in Figure 8, we can see that our method with $s = 3$ performs a little bit better than the others and that our method with $s = 2$ is equivalent to Spline IsoGeo and Spline L2.

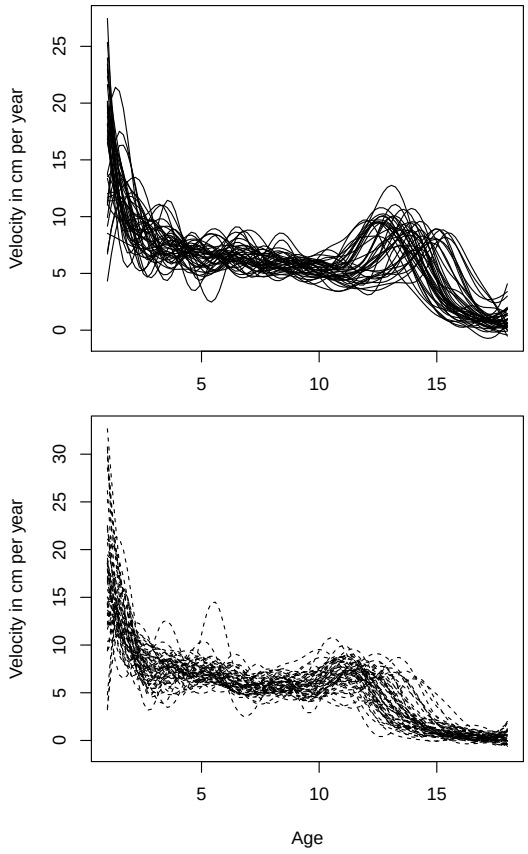

Figure 7: Velocity curves of 39 boys (top) and 54 girls (bottom).

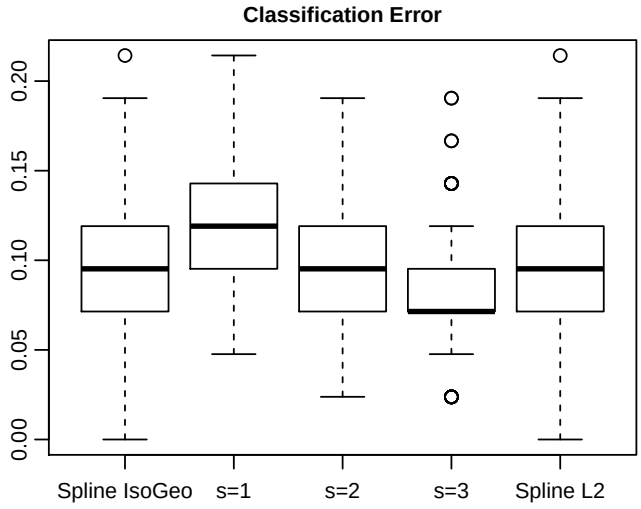

Figure 8: Classification errors for the velocity curves, our method with $s = 3$ performs better while our method with $s = 2$ is equivalent to Spline IsoGeo and Spline L2.

References

Chen, D., and Muller, H.-G. 2012. Nonlinear manifold representations for functional data. *The Annals of Statistics*

40(1):1–29.

Chen, Y.; Ho, S.; Freeman, P. E.; Genovese, C. R.; and Wasserman, L. 2015. Cosmic web reconstruction through density ridges: method and algorithm. *Monthly Notices of the Royal Astronomical Society* 454(1):1140–1156.

Diaconis, P.; Holmes, S.; and Shahshahani, M. 2013. *Sampling from a Manifold*, volume Volume 10 of *Collections*. Beachwood, Ohio, USA: Institute of Mathematical Statistics. 102–125.

Dijkstra, E. W. 1959. A note on two problems in connexion with graphs. *NUMERISCHE MATHEMATIK* 1(1):269–271.

Dimeglio, C.; Gallon, S.; Loubes, J.-M.; and Maza, E. 2014. A robust algorithm for template curve estimation based on manifold embedding. *Comput. Stat. Data Anal.* 70:373–386.

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

Floyd, R. W. 1962. Algorithm 97: Shortest path. *Commun. ACM* 5(6):345–.

Genovese, C. R.; Perone-Pacifico, M.; Verdinelli, I.; and Wasserman, L. 2014. Nonparametric ridge estimation. *The Annals of Statistics* 42(4):1511–1545.

Joshi, S.; Srivastava, A.; and Jermyn, I. 2007. Riemannian analysis of probability density functions with applications in vision. In *2007 IEEE Conference on Computer Vision and Pattern Recognition ; proceedings*. Piscataway, NJ: IEEE. 1664–1671.

Kneip, A., and Ramsay, J. O. 2008. Combining registration and fitting for functional models. *Journal of the American Statistical Association* 103(483):1155–1165.

Lin, B.; Yang, J.; He, X.; and Ye, J. 2014. Geodesic distance function learning via heat flow on vector fields. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, II–145–II–153. JMLR.org.

Ozertem, U., and Erdogmus, D. 2011. Locally defined principal curves and surfaces. *J. Mach. Learn. Res.* 12:1249–1286.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J.; Ramsay, J.; Silverman, B.; Silverman, H.; and Media, S. S. 2005. *Functional Data Analysis*. Springer Series in Statistics. Springer.

Ramsay, J. O.; Hooker, G.; and Graves, S. 2009. *Functional Data Analysis with R and MATLAB*. Springer Publishing Company, Incorporated, 1st edition.

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

Yeh, M. .; Lee, I. .; Wu, G.; Wu, Y.; and Chang, E. Y. 2005. Manifold learning, a promised land or work in progress? In *2005 IEEE International Conference on Multimedia and Expo*, 4 pp.–.