The hidden manifold distance for functional data

Susan Wei[1] **and Marie-Hélène Descary**[2]
[1] The University of Melbourne, [2]Université du Québec à Montréal

Abstract

Functional data analysis is the statistical analysis of smooth infinite-dimensional curves. The challenge of analyzing infinite-dimensional objects is ameliorated by the fact that the dimensionality of curves is only artifically high if they are smooth. There is also another sense in which the dimensionality of the curves under study may be low. This is the much less explored idea that the curves may actually live in a submanifold with low intrinsic dimension. Interesting classes of functional manifold data include classes of probability density functions and classes of warped curves of a common template function. In this work we address the estimation of the pairwise geodesic distance between functional manifold data when we only have access to their noisy realizations which live near, rather than exactly on, the manifold. The proposed methodology first projects the functional data on to the underlying hidden manifold and then performs operations on this hidden manifold. Good estimation of the pairwise geodesic distance has beneficial implications for many downstream tasks in functional data analysis which we illustrate in the case of distanced-based functional classification.

Introduction

Many statistical methods depend on a measure of distance. Methods in functional data analysis, the study of infinite-dimensional smooth curves, are no exception. Certain classes of functional data may naturally live on manifold, e.g. classes of probability density functions and classes of warped curves of a common template function. In situations where the manifold hypothesis might be plausible, one might consider using a geodesic distance that takes into account the intrinsic structure of the manifold rather than the standard L^2 distance. The benefits of adopting the geodesic distance may be realized in downstream tasks such as distanced-based clustering and classification of functional obvervations.

Specifically, let $X_1, \dots, X_n$ be a sample of n independant 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 low-dimensional nonlinear Riemannian manifold $\mathcal{M} \subset L^2([a,b],\mathbb{R})$ where g is a Riemannian metric tensor on $\mathcal{M}$ which can be 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

$$\begin{aligned} d_g(X_i,X_j) \quad &:= \quad \inf\big\{l(\gamma) : \gamma : [a,b] \to \mathcal{M} \text{ piecewise smooth}, \\ &\qquad \gamma(a) = X_i, \gamma(b) = X_j]\big\}, \end{aligned}$$

where

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

is the length of a smooth curve $\gamma : [a,b] \subset \mathbb{R} \to \mathcal{M}$.

Given perfect functional observations X_i and X_j, i.e. with no measurement error and on a very dense domain grid, the estimation of the geodesic distance $d_g(X_i,X_j)$ can be accomplished straightforwardly using a shortest-path algorithm, e.g. the Floyd-Warhsall algorithm (Floyd 1962). However, when functional data are observed with noise, this approach is likely to fail. This is because the error may be such that the noisy functional manifold data no longer live on a manifold. A preprocessing step where the nosiy functional manifold data is smoothed is also likely to fail. This is because smoothing techniques in functional data analysis are designed for the L_2 recovery of the original curve, not to ensure the recovered functional versions of the data lie on or close to the functional manifold $\mathcal{M}$.

In this work, we put forth a technique for the specific task of recovering the $n \times n$ matrix G of pairwise geodesic distances, i.e.

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

when we have, in lieu of X_i and X_j, access only to discretely-observed noisy functional observations Y_i and Y_j that possibly live off the true manifold. This is accomplished by first projecting the functional data on to the underlying hidden manifold and then performing operations on this hidden manifold.

Related work

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. Of particular relevance to our work is the modified ISOMAP procedure, called P-ISOMAP, introduced in Chen and Muller (2012), which adds a data-adaptive penalty to allow for noisy functional observations whereas in the classic ISOMAP algorithm, observations are assumed to lie exactly on the manifold. Another existing work for estimating the pairwise geodesic distance for functional manifold data can be found in Dimeglio et al. (2014). The estimator is called Robust-ISOMAP for the fact that it is motivated by modifying the ISOMAP algorithm to be less sensitive to outliers. We will discuss both P-ISOMAP and robust ISOMAP in more details in the simulation section.

Proposed method for estimating geodesic distances

Suppose that each curve X_i is observed with measurements errors on a grid $T_i = (t_{i1}, \dots, t_{iK})$, i.e. we observe a sample of K-dimensional vectors $Y_1, \dots, Y_n$ with $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$, where the random variables ϵ_{ij} have mean zero and are uncorrelated with each other. We assume that each observation grid $T_1, \dots, T_n$ is dense.

We begin by converting the discrete functional observations into continuous ones. Let $\tilde{X}_1, \dots, \tilde{X}_n$ denote the functional versions of the raw data obtained by some smoothing method. Our proposed methodology for estimating the pairwise geodesic distances $\{d_{\mathcal{M}}(X_i, X_j)\}_{i>j}$ is agnostic to the smoothing method employed. For example we may employ spline smoothing 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 $\tilde{X}_i$.

Before introducing our procedure, let us first review the ISOMAP algorithm (Tenenbaum, Silva, and Langford 2000), a three-step procedure that takes a set of points $x_1, \dots, x_n$ in $(\mathcal{M}, g)$, a submanifold of $\mathbb{R}^D$ as input and produces an embedding of the input data in the space $\mathbb{R}^d$ with $d < D$ that preserves pairwise geodesic distances.

1. Construct a weighted graph G with nodes corresponding to the observations $x_1, \dots, x_n \in \mathbb{R}^D$. Two nodes x_i and x_j are connected by an edge e_{ij} if the distance $d_{ij} = ||x_i - x_j||_2$ is smaller than a given value ϵ, and the weight associated to an edge e_{ij} is d_{ij}.
2. Estimate the pairwise geodesic distances $d_g(x_i, x_j)$ based on G using shortest-path algorithms. Specifically, the geodesic distance between two nodes is estimated to be the length of the shortest path in G between these two nodes, i.e./ 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 weigthed graph but they do not produce the weighted graphs themselves. Formulating the graph in step 1 of ISOMAP is in fact the most challenging asepct of the algorithm since it depends heavily on the choice of the tuning parameter ϵ. 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.

In what follows, we call IsoGeo the procedure which performs a modified version of the first step of ISOMAP, proposed in @Dimeglio2014, followed by the original second step of ISOMAP. The first step of IsoGeo modifies Step 1 of ISOMAP so that the constructed graph G is independent of the tuning parameter ϵ: 1'. Construct the complete weigthed graph G_c with nodes corresponding to the observations $x_1, \dots, x_n \in \mathbb{R}^D$. Obtain the minimal spanning tree G_s associated with G_c and denote its set of edges by E_s. The graph G is obtained by adding all edges e_{ij} to E_s for which the following condition is true:

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

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

Our method is based on the idea that the underlying functional manifold can be sufficiently well-recovered by the subspace-constrained mean-shift algorithm (Ozertem and Erdogmus 2011). Theoretical justificaton for this can be found in Genovese et al. (2014). We expect the points $\tilde{X}_1^{s,\hat{\mathcal{M}}}, \dots, \tilde{X}_n^{s,\hat{\mathcal{M}}}$ to lie close to the real manifold $\mathcal{M}$ and then $d_{\hat{\mathcal{M}}}$ to be close to $d_{\mathcal{M}}$. Our procedure is described by the following steps:

1. Discretise the functions $\tilde{X}_1, \dots, \tilde{X}_n$ by evaluating them on a common grid $\tilde{T} = (t_1, \dots, t_K)$. Let s be a positive integer, smaller than K. Obtain $\tilde{X}_1^s, \dots, \tilde{X}_n^s \in \mathbb{R}^s$ using multidimensional scaling whereby the pairwise $\mathbb{R}^K$ Euclidean distances on the discretised functions $\tilde{X}_i$ are preserved.
2. Apply the subspace constrained mean-shift algorithm (Ozertem and Erdogmus 2011) to each of $\tilde{X}_i^s$ to obtain $\tilde{X}_i^{s,\hat{\mathcal{M}}}$.
3. Use IsoGeo to approximate the pairwise geodesic distances $\{d_{\hat{\mathcal{M}}}(\tilde{X}_i^{s,\hat{\mathcal{M}}}, \tilde{X}_j^{s,\hat{\mathcal{M}}})\}_{i>j}$.

Thus our geodesic distance estimator is the $n \times n$ matrix $\hat{G}$ whose elements are given by

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

Since ridge estimation suffers from the curse of dimensionality, we first reduce the dimension of our data with multidimensional scaling in Step 1 before applying the mean-shift algorithm.

Selection of tuning parameters

Describe heuristics. We pick the bandwidth h in the subspace constrained mean-shift using Equation (A1) of (Chen et al. 2015).

Simulation study

We perform a simulation study to ascertain the efficacy of our method for estimating pairwise geodesic distances when one only has access to discretely-observed noisy functional data. Three different metrics are used to assess the quality of a pairwise geodesic distance estimator.

- the near isometry metric: this is the area under the receiver operating curve with ϵ on the x-axis and the degree to which near-ϵ isometry holds on the y-axis. Near-ϵ isometry is a relaxation of isometry measured by the percentage of estimated pairwise distances between $1-\epsilon$ and $1+\epsilon$ of the truth pairwise distance.
- the relative Frobenius metric: this is given by $||d-\hat{d}||/||d||$ where d and $\hat{d}$ are the true and estimated $n \times n$ geodesic distance matrices, respectively.
- the Pearson correlation metric: this is the Pearson correlation coefficient between the upper diagonal of d and $\hat{d}$.

Alternative estimators of geodesic distance [Marie]

We compare the performance of our method to the one of the three following methods:

1. *Raw Data (RD)* The naive approach consisting in appling directly IsoGeo on the raw vectors $Y_1, \ldots, Y_n \in R^K$. ¡!– to obtain an estimator $\hat{G}_{\text{RD}}$.–¿ Note that this procedure only makes sense if all the grids $T_1, \ldots, T_n$ are the same.
2. *Spline Smoothing (SS)* The natural method consisting in applying IsoGeo on the smoothed version $\tilde{X}_1, \ldots, \tilde{X}_n$ of the raw data obtained by spline smoothing as described in (1).
3. *P-ISOMAP (pI)* The two step procedure developed in @ChenMuller2012 where the first step is a modified version of step I of ISOMAP incorporating a penalty in order to robustify the construction of the weighted graph and the second step is step II of ISOMAP.

We ran Chen and Muller's P-ISOMAP separately in Matlab with their automatic penalty selection and discovered that ???.

Simulation scenarios

Isometric functional manifold of normal density functions This scenario is modified from what is referred to as Manifold 2 in (Chen and Muller 2012) by fixing the variance of the normal density to be 1. We have

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

with the L_2 inner product as the metric tensor of $\mathcal{M}$, where $X_\beta : [a,b] \to \mathbb{R}$ is given by $X_\beta(t) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(t-\beta)^2\right]$. We set $a=-4$ and $b=4$.The geodesic distance between the curves X_{β_1} and X_{β_2} is given by

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

where the approximation comes from the fact that we are integrating on $[a,b] = [-4,4]$ and not on $\mathbb{R}$. We can see this manifold is isometric, since the geodesic distance between X_{β_1} and X_{β_2} in $\mathcal{M}$ is the Euclidan distance between the β's, up to some scaling factor. Note that the "straight" line connecting X_{β_1} and X_{β_2} in $\mathcal{M}$ does not always stay inside of $\mathcal{M}$, so we cannot employ the calculation technique of Scenario 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 has a nice closed form geodesic. 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 metric tensor given by 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 concides with the $L_2[0,1]$ inner product. (Joshi, Srivastava, and Jermyn 2007) showed that the geodesic distance between any two ψ_1 and ψ_2 in $\mathcal{M}$ is simply

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

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

$$M = \{\psi_{\alpha,\beta} : 1 \leq \alpha \leq 5, 2 \leq \beta \leq 5\}$$

where $\psi_{\alpha,\beta} : [0,1] \to \mathbb{R}$ is the pdf of $Beta(\alpha, \beta)$.

Functional manifold of warping functions This is based on Equations (17) and (18) of(Kneip and Ramsay 2008) but with z_{i1}, z_{i2} set to 1. (Equation 17 has a typo where the exponentials are missing negative signs).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

$$h_\alpha(t) = 6\frac{\exp\{\alpha(t+3)/6\}-1}{\exp\{\alpha\}-1}, \alpha \neq 0$$

and $h_\alpha(t) = t$ if $\alpha = 0$. Consider the manifold

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

. The geodesic distance is then

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

Sampling from a functional manifold

As of yet, there is little work as to how to sample from a functional manifold. Even in the Euclidean case, it is not obvious how sampling should be done (Diaconis, Holmes, and Shahshahani 2013). To safeguard against sampling unevenly on the functional manifold, we sample on a very concentrated measure for the intrinsic parameters. For both the manifold of normal densities and the manifold of warping functions, we sample α according to ???For the manifold of square root beta densities, we sample as follows How to sample properly from a functional manifold could be interesting future work.

Results

The cost of employing proposed method when manifold is flat

Distance-based functional classification [Marie]

In this section, we explore whether our geodesic distance estimator has benefits for downstream analysis task. There are many tasks we could consider here such as distance-based nonparametric regression and distanced-based functional clustering, but we will focus on distance-based functional classification. It must be noted that while curve alignment, also known as curve registration, is necessarily performed as a preprocessing technique prior to clustering and classification, our geodesic distance estimator allows one to forsake this step. For simplicity, assume the task is binary classification. Associated to each functional object x is a binary y indicating class membership. Consider the classifier proposed in Ferraty and Vieu (2003,2006) which is a functional version of the Nadaraya-Watson kernel estimator of class membership probabilities:

$$\hat{p}(y=0|x)\frac{\sum_{i=1}^n K[h^{-1}d(x,x_i)]1(y_i=0)}{\sum_{i=1}^n K[h^{-1}d(x,x_i)]}$$

We shall compare our method to using L_2 distance, possibly weighted, and with curve registration already accomplished. Describe alternative methods in detail.

The bandwidth in the classifier should be tuned individually for each method. Also we might need to tune MDS dimension s since in real data, the dimension of the manifold might be much higher than encountered in the simulation scenarios where it never goes above 2.

Datasets used by functional classification papers

- Wheat, rainfall and phoneme in Aurore's paper "Achieving near-perfect classification for functional data"
- Berkeley growth curves in (Chen, Reiss, and Tarpey 2014).
- Tecator and phoneme in (Galeano, Joseph, and Lillo 2015) Mahalanobis technometrics paper.
- yeast cell cycle gene expression (can't find this publicly) in (Leng and Muller 2005) "Classification using functional data analysis for temporal gene expression data"

Datasets used in functional manifold papers

- Berkeley growth, yeast cell cycle gene expression (can't find this publicly) in (Chen and Muller 2012)
- Tecator in (Lin and Yao 2017) contamination paper
- Berkeley growth, gait cycle in (Dimeglio et al. 2014) robust isomap paper

References

[Chen and Muller 2012] Chen, D., and Muller, H.-G. 2012. Nonlinear manifold representations for functional data. *The Annals of Statistics* 40(1):1–29.

[Chen et al. 2015] 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.

[Chen, Reiss, and Tarpey 2014] Chen, H.; Reiss, P. T.; and Tarpey, T. 2014. Optimally weighted l2 distance for functional data. *Biometrics* 70(3):516–525.

[Diaconis, Holmes, and Shahshahani 2013] 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 1959] Dijkstra, E. W. 1959. A note on two problems in connexion with graphs. *NUMERISCHE MATHEMATIK* 1(1):269–271.

[Dimeglio et al. 2014] 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.

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

[Galeano, Joseph, and Lillo 2015] Galeano, P.; Joseph, E.; and Lillo, R. E. 2015. The mahalanobis distance for functional data with applications to classification. *Technometrics* 57(2):281–291.

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

[Joshi, Srivastava, and Jermyn 2007] 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 and Ramsay 2008] Kneip, A., and Ramsay, J. O. 2008. Combining registration and fitting for functional models. *Journal of the American Statistical Association* 103(483):1155–1165.

[Leng and Muller 2005] Leng, X., and Muller, H.-G. 2005. Classification using functional data analysis for temporal gene expression data. *Bioinformatics* 22(1):68–76.

[Lin and Yao 2017] Lin, Z., and Yao, F. 2017. Functional Regression on Manifold with Contamination. *arXiv e-prints* arXiv:1704.03005.

[Lin et al. 2014] 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 and Erdogmus 2011] Ozertem, U., and Erdogmus, D. 2011. Locally defined principal curves and surfaces. *J. Mach. Learn. Res.* 12:1249–1286.

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