Exploring different semi-metrics in regression and classification of functional data.

Alejandra Cabaña
based on joint work with Amanda Fernández and Oriol Datzira

Universitat Autònoma de Barcelona

ECODEP Seminar, October 19, 2022

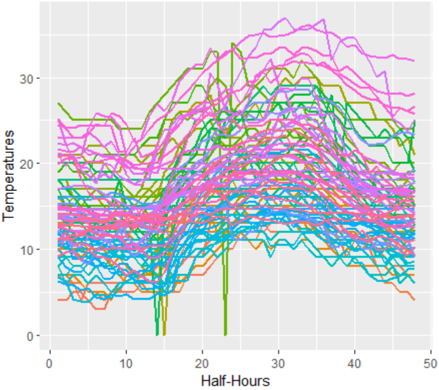
Random curves

- In many situations coming from different fields of applied sciences the collected data can be considered as curves.
- We call functional data to a collection of observations of a random process that is registered as a function, $X_i : I \subset \mathbb{R} \to \mathbb{R}^n$, although the domain need not be a subset of $\mathbb{R}$.
- In practice, the concept of a continuum measurement doesn't exist. Processes are monitored in a discrete grid $t_1, t_2, ..., t_N$. So at the end, one always has a vector observation $(x(t_1), ..., x(t_N))$.
- There are typically two reasons to consider this as functional data:

1. the theoretical possibility of observing the phenomenon in a much finer grid,
2. and the choice of a functional model to approximately represent it.

Functional data.

A random variable $\mathcal{X}$ is called a **functional variable** if it takes values in an infinite dimensional space (or functional space). An observation x of $\mathcal{X}$ is called a functional observation.



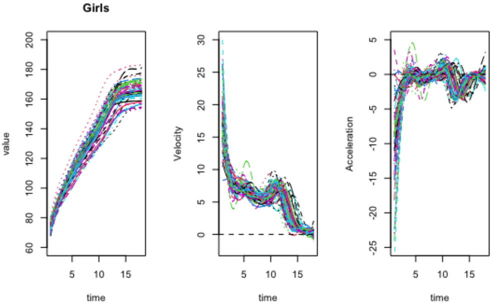
Half-Hourly temperatures measured at Adelaide airport from Saturday to Sunday between 6/7/1997 and 31/3/2007.

Closeness for functional data.

Proximity measures between mathematical objects play a major role in all statistical methods.

In many situations, a classical norm can be used to measure the closeness between functional objects.

However, some situations, it appears that semi-metric spaces are better adapted than metric spaces.



Semi-metrics.

Proximities between functional variables are measured using semi-norms and semi-metrics.
$||\cdot||$ is a semi-norm if it satisfies:

1. $\forall(\lambda, x) \in \mathbb{R} \times E, \quad ||\lambda x|| = |\lambda| ||x||,$
2. $\forall(x, y) \in E \times E, \quad ||x+y|| \leq ||x|| + ||y||.$

Let E be a functional space. A semi-metric d in E is a function:

$$\begin{aligned} d:\ & E \times E \rightarrow \mathbb{R}^+ \\ & (x, y) \mapsto d(x, y) = ||x - y|| \end{aligned}$$

satisfying:

1. $\forall x \in E, \quad d(x, x) = 0,$
2. $\forall(x, y, z) \in E \times E \times E, \quad d(x, y) \leq d(x, z) + d(z, y).$

A semi-metric is similar to a metric with the exception that $d(x, y) = 0$ not necessarily implies $x = y$.

Nonparametric regression for functional data.

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ be n observations of p predictors and let Y_i be the response continuous variables for $i = 1, \ldots, n$ i.i.d. as $(\mathbf{X}, Y)$. For a new observation $\mathbf{x} = (x_1, \ldots, x_p)$ the regression operator is defined as:

$$r(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}),$$

and can be consistently estimated with[1]

$$\hat{y} = \widehat{r_n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K(w_1 d_1(X_{i1}, x_1) + \cdots + w_p d_p(X_{ip}, x_p))}{\sum_{i=1}^{n} K(w_1 d_1(X_{i1}, x_1) + \cdots + w_p d_p(X_{ip}, x_p))},$$

where K is a kernel, d_i are semi-metrics, and $w_1, \ldots, w_p$ positive weights computed from the observed data working as smoothing parameters (essentially, inverses of bandwidths).
Observe that it's an extension of Nadaraya-Watson density estimator.

---

[1] Ferraty & Vieu (2006), *Nonparametric Functional Data Analysis: Theory and Practice.*

Regularity conditions[2]

The above estimator satisfies

$$\sum_n P(|\widehat{r_n}(\mathbf{x}) - r(\mathbf{x})| > \epsilon) < \infty,$$

provided

- $r \in \mathcal{C}^0 = \left\{f : E \to \mathbb{R} \forall X, X' \lim_{d(X,X')\to 0} f(X) = f(X')\right\}$, which ensures that the regression operator r is continuous.
- $P(d(\mathcal{X}, x) < \epsilon) = \mathbb{E}(1_{B(x,\epsilon)}(\mathcal{X})) = \gamma(\epsilon) > 0 \quad \forall \epsilon > 0,$
- and the bandwidths $h_i = w_i^{-1}$ satisfy $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} \frac{\log n}{n\gamma(h_n)} = 0$.
- The response variables Y satisfies, for all $m \geq 2$, $\mathbb{E}(|Y^m||\mathcal{X} = x) < \sigma_m(x) < \infty$, with $\sigma(\cdot)$ continuous at x.

This implies that $\widehat{r_n}(\mathbf{x})$ converges both in probability and almost surely to $r(\mathbf{x})$.

---

[2] Ferraty-Vieu (2006)

Nonparametric classification.

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ be n observations of p predictors and let Y_i the categorical response variables ($Y_i \in \mathcal{G} = \{1, \ldots, G\}$) for $i = 1, \ldots, n$ i.i.d. as $(\mathbf{X}, Y)$. For a new observation $\mathbf{x} = (x_1, \ldots x_p)$ the posterior probabilities to predict the class where $\mathbf{x}$ belongs is defined as:

$$P_g(\mathbf{x}) = \mathbb{E}\left(1_{[Y=g]} | \mathbf{X} = \mathbf{x}\right),$$

which, as before, can be estimated with

$$\hat{P}_g(\mathbf{x}) = \frac{\sum_{i=1}^n 1_{[Y_i=g]} K(w_1 d_1(X_{i1}, x_1) + \cdots + w_p d_p(X_{ip}, x_p))}{\sum_{i=1}^n K(w_1 d_1(X_{i1}, x_1) + \cdots + w_p d_p(X_{ip}, x_p))},$$

and

$$\hat{y} = \arg\max_{g \in \mathcal{G}} \hat{P}_g(\mathbf{x}).$$

Conditions for consistency of the estimator are similar to the previous ones.

The semi-metrics we have used.

1. **L^2 distance:**

$$d_2(X,Y) = ||X-Y||_2 = \left(\int |X(t) - Y(t)|^2 dt \right)^{\frac{1}{2}}.$$

2. **Hausdorff distance:**

$$d_H(X,Y) = max \left\{ \underset{x \in X}{sup} \left\{ \underset{y \in Y}{inf} d(x,y) \right\}, \underset{y \in Y}{sup} \left\{ \underset{x \in X}{inf} d(y,x) \right\} \right\}.$$

3. **Wasserstein distance:**

$$d_{W,p}(P,Q) = \left(\underset{J \in \mathcal{J}(P,Q)}{inf} \int ||x-y||^p dJ(x,y) \right)^{1/p},$$

where P and Q are probability measures defined in a metric space (M,d) with finite p-th moments, and $\mathcal{J}$ is the set of probability measures J with marginals P, Q.

BirdChicken Data Set[3]

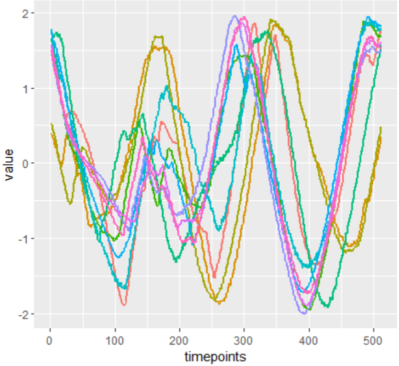
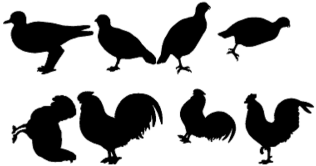


Figure: Representation of some observations from the training set of the BirdChicken data set.

Goal: classify functional data consisting on distances from center to contour of images.



Accuracy

L^2	Hausdorff	Wasserstein
0.45	0.65	0.9

[3] `https://timeseriesclassification.com/description.php?Dataset=BirdChicken`

ArabicDigits Data Set[4]

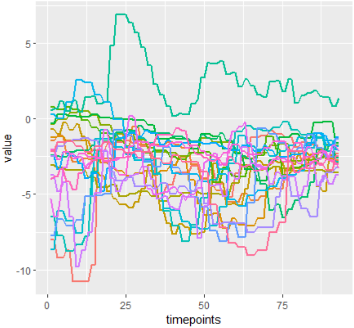


Figure: Representations of some time series of MFCC2.

Goal: Multiclass classification.

8800(10 digits x 10 repetitions x 88 speakers) time series of 13 Frequency Cepstral Coefficients (MFCCs) Arabic native speakers between the ages 18 and 40 to represent ten spoken Arabic digit. Due to computational issues only with 3 classes and 4 predictors were used.

Accuracy

L^2	Hausdorff	Wasserstein
0.978979	0.6846847	0.8348348

[4] `https://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit`

Tecator Data Set[5]

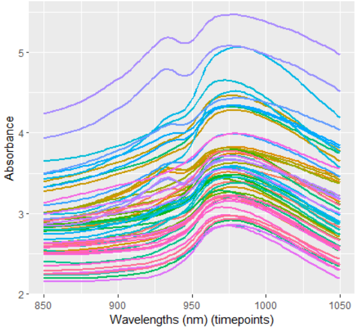


Figure: Spectrometric curves of finely chopped meat.

GOAL: Predict scalar response, functional and scalar covariates.

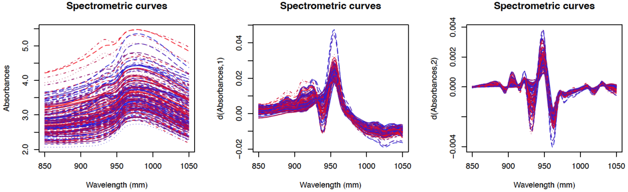
Values of spectrometric curves of 215 pieces of finely chopped meat, corresponding to the absorbance measured at 100 wavelengths (from 850 mm to 1050 mm) and the percentages of fat, water and protein of each piece determined by analytic chemistry. The objective is to predict the percentage of Fat of a piece, knowing the values of absorbance and the percentage of Water and Protein.
The data set contains 3 predictors ($p = 3$) in order to predict the quantity of Fat of a new piece of chopped meat, where the spectrometric curves is a functional one and the water and protein of the piece are scalar predictors.

	MSE	
L^2	Hausdorff	Wasserstein
2.055423	1.963038	1.963302

[5] Ferraty & Vieu, `fda.usc` R library

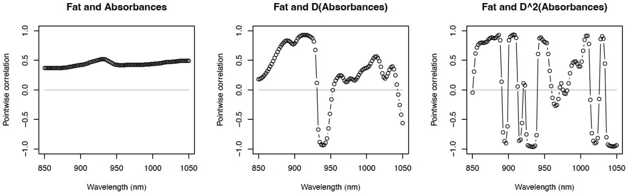
More information that could have been used...

Spectrometric curves and their derivatives



Low fat content High fat content

Pointwise Correlations between fat and absorbances



Comments and further things to do.

- Performance varies broadly using different (semi-)metrics.
- Computational burden makes L^2 differs widely as well.
- Explore the use of semi-metrics that include derivatives of the curves.
- Haussdorff distance has successfully been used with highly nondifferenciable curves[6].
- If the curves are smooth, the L^2 distance would be probably the best option.
- A further research problem could be to see whether, when dealing with rough curves, a previous smoothing (splines, for instance) leads to an improvement of the performance of the L^2 distance based methods.

[6] Fernández-Fontelo *et.al.(2022)* `https://arxiv.org/abs/2205.13380`