Nonparametric Classification and Regression for Functional Data

Oriol Datzira, Alejandra Cabaña, Amanda Fernández
Oriol.Datzira@autonoma.cat

Universitat Autònoma de Barcelona

Problem

Let $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$, for $i = 1, \dots, n$, be n observations of p functional, continuous or categorical predictors of a certain phenomenon and let Y_i for $i = 1, \dots, n$ be the n response continuous or categorical variables. Hence, we can consider the pairs $(\mathbf{X}_i, Y_i)_{i=1,\dots,n}$ i.i.d.
For a new observation $\mathbf{x} = (x_1, \dots, x_p)$ of the same phenomenon, a nonparametric prediction based on a kernel of the unknown response variable y of $\mathbf{x}$ is given by:

$$\hat{y} = \hat{r}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K(w_1 d_1(X_{i1}, x_1) + \dots + w_p d_p(X_{ip}, x_p))}{\sum_{i=1}^{n} K(w_1 d_1(X_{i1}, x_1) + \dots + w_p d_p(X_{ip}, x_p))} \quad if \;\; Y_i \in \mathbb{R} \quad \text{i.e. } \textbf{Regression Tasks}$$

or

$$\hat{y} = \underset{g \in \mathcal{G}}{argmax}\, \hat{P}_g(\mathbf{x}) = \frac{\sum_{i=1}^{n} 1_{[Y_i = g]} K(w_1 d_1(X_{i1}, x_1) + \dots + w_p d_p(X_{ip}, x_p))}{\sum_{i=1}^{n} K(w_1 d_1(X_{i1}, x_1) + \dots + w_p d_p(X_{ip}, x_p))} \quad if \;\; Y_i \in \mathcal{G} = \{1, \dots, G\} \quad \text{i.e. } \textbf{Classification Tasks}$$

with K being a kernel, d a semi-metric and $w_1, \dots, w_p$ positive weights computed from the observed data working as smoothing parameters.

We propose different semi-metrics to test the performance of the above classification/regression methods for real-data problems.

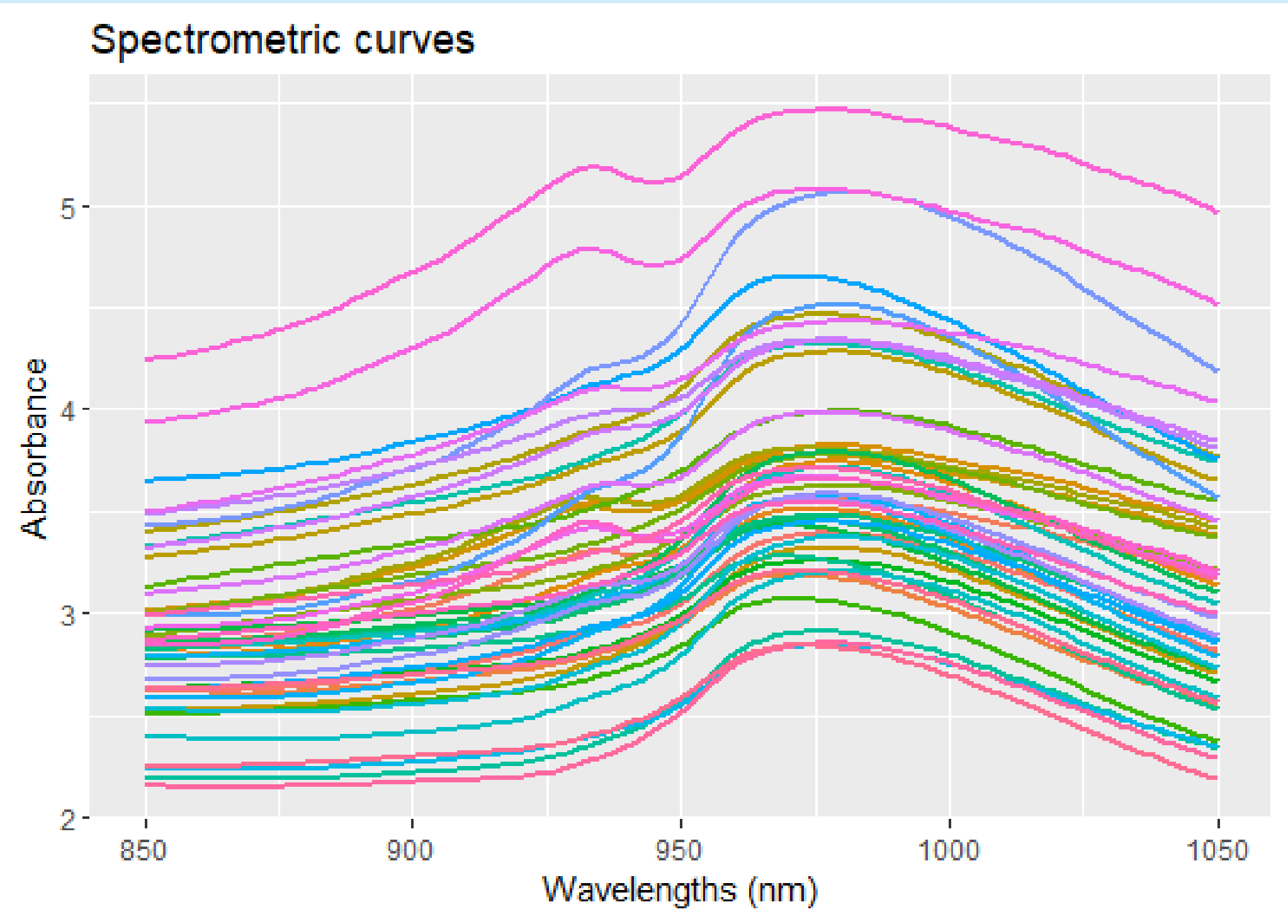
L^2 distance

The L^2 is the classical Euclidian norm for functional objects.
Let E be a functional space. Let X, Y be two functional elements of E. The L^2 distance d_2 between X and Y, $d_2(X, Y)$ is defined as:

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

Tecator

For each peace of finely chopped meat, the *Tecator* data set contains the values of spectrometric curves of 215 peaces which corresponds to the absorbance measured at 100 wavelengths (from 850mm to 1050mm) and the percentages of fat, water and protein of each peace determined by analytic chemistry and **the objective is to predict the percentage of Fat of a piece knowing the values of absorbance and the percentage of Water and Protein** .



As we can see, the functional curves of the *Tecator* data set are smooth so it seems, a priori, that the L^2 distance will make a better performance. However, its slightly poorer than the other two, which are equivalent . The performance in terms of the Mean Squared Error are:

	L^2	Hausdorff	Wasserstein
MSE	2.055423	1.963038	1.963302

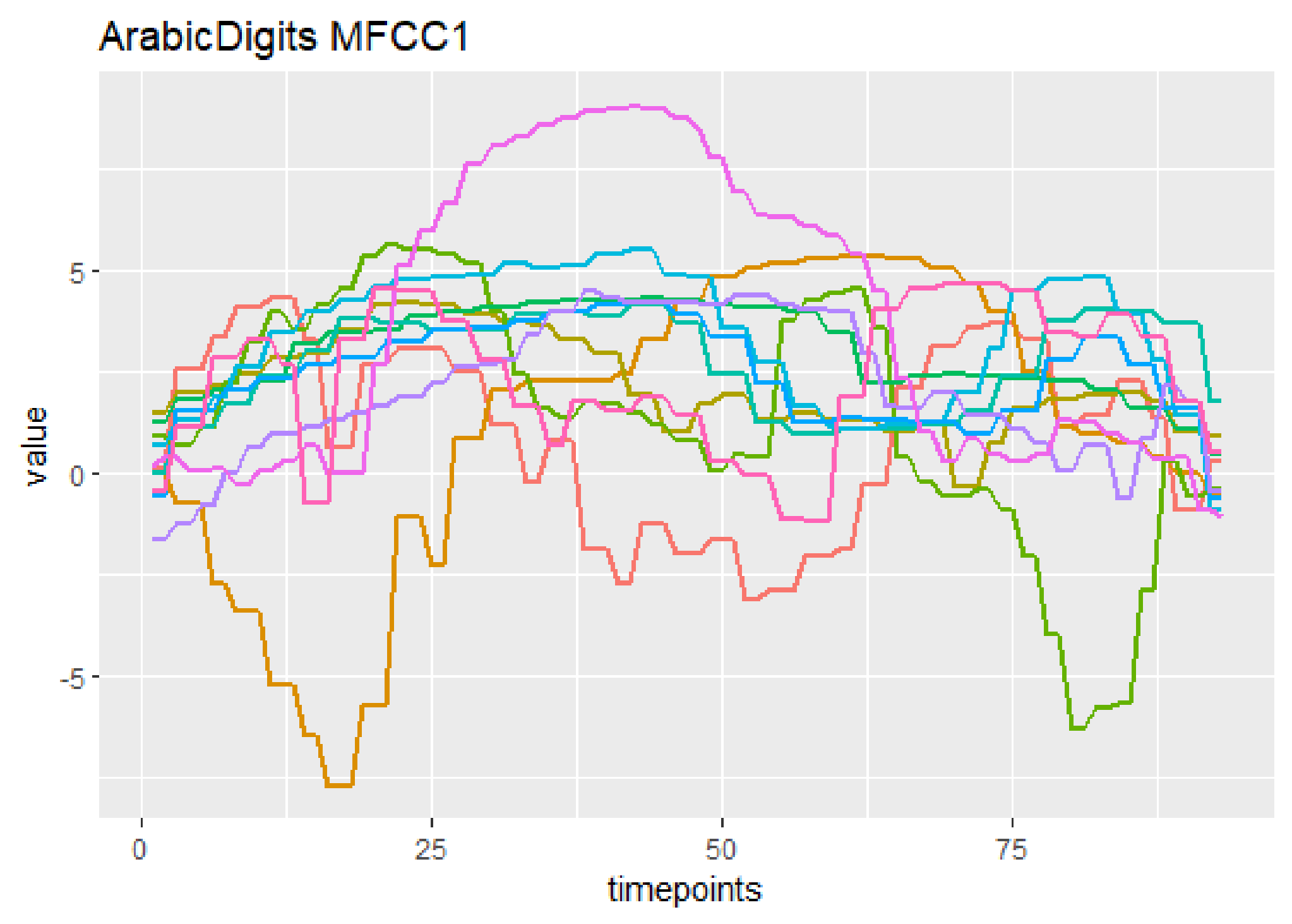
Hausdorff distance

The Hausdorff Distance is a measure of dissimilarity between two point sets.
Let X and Y be two sets $\neq \emptyset$ of a (semi-)metric space (E, d). Their Hausdorff distance $d_H(X, Y)$ can be defined as:

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

ArabicDigits

The *ArabicDigits* data set contains time series of 13 Mel Frequency Cepstrum Coefficients (MFCCs) which correspond to 10 spoken arabic digits. **The objective is to be able to classify the spoken arabic digit knowing the 13 MFCCs.**



To test the performance we have used 4 predictors (4 MFCCs), as well as just 3 labels (3 spoken arabic digits) into which classify and just a part of the entire data set.
In this case, since it seems important to compute the proximities between predictors preserving the order of the timepoints, the L^2 distance seems to be the best (semi-)metric to pick because the way how it is computed, contrary to how the Hausdorff distance computes proximities. The performance using the (semi-)metrics in terms of the accuracy are:

	L^2	Hausdorff	Wasserstein
Acc.	0.978979	0.6846847	0.8348348

Wasserstein distance

The Wasserstein Distance arises from the idea of *optimal transport.*
Let $T : \mathbb{R}^q \to \mathbb{R}^q$ and $X \in \mathbb{R}^q$. The distribution of $T(X)$ is known as the push-forward of P:

$$T_{\#} P(A) = P(\{x : T(x) \in A\}) = P(T^{-1}(A)).$$

The optimal transport distance is defined as

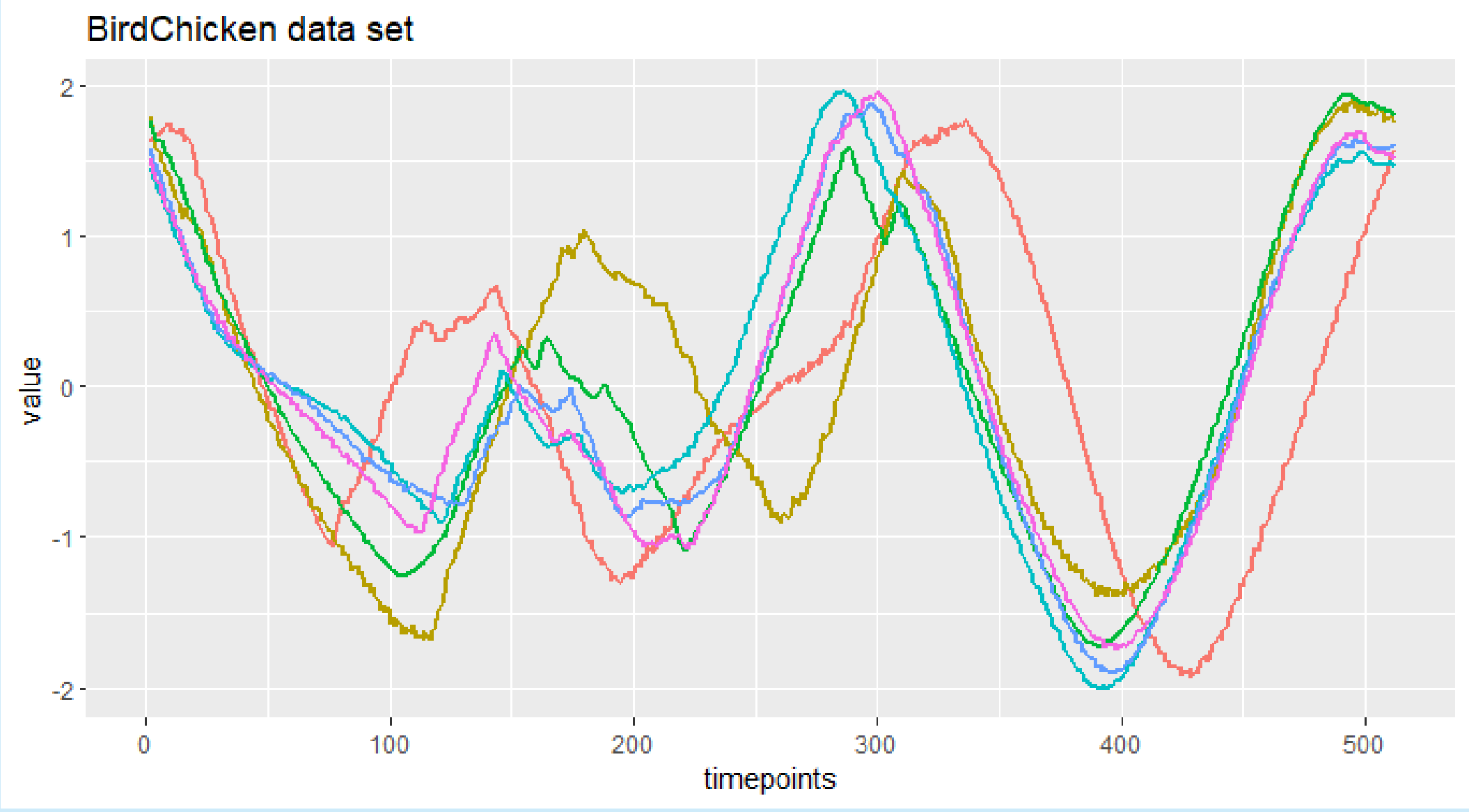
$$\underset{T}{inf} \int ||x - T(x)||^p dP(x),$$

where the infimum is over all T that satisfies $T_{\#} P = Q$. It measures how far you have to move the mass of P to turn it into Q.
Let $\mathcal{J}(P, Q)$ denote all the joint distributions J for (X, Y) that satisfy that $T_{X\#} J = P$ and $T_{Y\#} J = Q$, where $T_X(x, y) = x$ and $T_Y(x, y) = y$. Then, the Wasserstein distance $d_{W,p}$ for $p \geq 1$ between P and Q is defined as:

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

BirdChicken

The *BirdChicken* data set is a database that consists of the 1 dimensional series of distances to the centre of the outlines of contour images of chicken and birds. **The aim of the Bird/Chicken problem is to be able to distinguish between an outline of a bird and a chicken**.



In this case, as the functional data are non-smooth, we expect a bad performance of the L^2 distance and since the type of problem is to distinguish contours of images, which is similar to the optimal transport problem, the Wasserstein distance seems, a priori, to be the best pick. The performance of the (semi-)metrics in terms of the accuracy are:

	L^2	Hausdorff	Wasserstein
Acc.	0.45	0.65	0.9

Conclusions

We have tested nonparametrical methods for functional and/or categorical/continuous data using different (semi-)metrics and, as expected, the performance varies broadly depending on the distance used, which implies that, to get a good performance, a previous study of the type of data in hand, as well as the type of problem, is crucial in order to pick the (semi-)metrics that fits better to the data set. L^2 is the less computationally expensive, of course.