Machine Learning with Functional Data

Exploratory analysis and alignment

Depth measures

The study of *depth measures* has become a recurrent topic in FDA.

Given a probability measure P on the sample space $\mathcal{X}$, a *depth function* $D(P, x)$ is a non-negative function defined on $\mathcal{X}$ that indicates how deep an observation x is in the distribution P. Let $\mathbb{P}_n$ be the empirical distribution of the sample $X_1, \ldots, X_n$, then $D(\mathbb{P}_n, x)$ indicates the depth of x in that sample.

For univariate data, typical depth functions include the mean, median, and mode. For multivariate data, computing depth measures becomes increasingly difficult as the dimension increases.

Desirable properties of a depth measure include: affine invariance, maximality at the center, monotonicity, zero at infinity, and invariance under dimensionality reduction.

Given a depth measure, one can define the median as the deepest point in the sample. Similarly, percentiles and truncated means can be defined.

Fraiman-Muniz Depth is an integral depth measure based on the naturalness of order statistics in one dimension.

Let $F_{n,t}$ be the empirical distribution function of $x_1(t), \ldots, x_n(t)$. The univariate depth of each data point $x_i(t)$ is denoted by $D_i(t) = 1 - |\frac{1}{2} - F_{n,t}(x_i(t))|$. The depth index for each i is defined as follows:

$$I_i = \int_0^1 D_i(t)dt.$$

Integrated Depth: A general approach to get an infinite-dimensional depth function from the one-dimensional depths of the projections is proposed Cuevas and Fraiman (2009): the basic idea is just to define a new depth function by integrating out the one-dimensional depths. This leads to a definition of type

$$D(P, x) = \int D_i(P_f, f(x)) dQ(f),$$

where D_i is a one-dimensional depth function, f denotes an element in the continuous dual space, P_f is the (one-dimensional) distribution of $f(X)$ (the projection of X via the real linear continuous f) and Q is a probability measure on the continuous dual space of f's.

the methods of class IntegratedDepth can be used to compute integrated depth measures. The default choice is the measure propose by Fraiman and Muniz.

The practical aspects of this idea, as well as some comparisons with other methods, have been considered in Cuevas, Febrero and Fraiman (2007).
We shall not discuss here depth functions in further detail. Let us just remark the fact that every depth function has an associated notion of median (the deepest point), quantiles (defined in the obvious way after sorting the points according to their depths) and α-trimmed mean

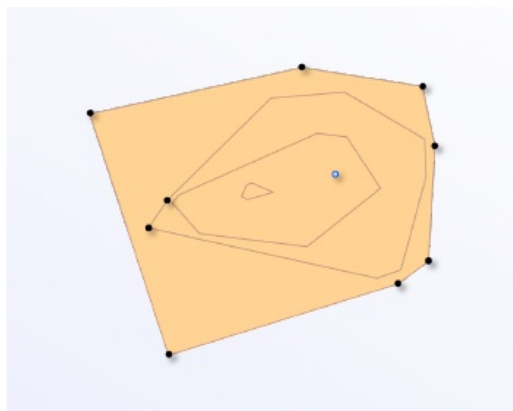
Tukey depth

In the case $\mathcal{X} = \mathbb{R}$, the usual depth functions (which are equivalent in various cases) are given by

$$D_0(P, x) = P(-\infty, x]P[x, \infty) \text{ and } D_1(P, x) = \min(P[x, \infty), P(-\infty, x]).$$

Tukey depth: In both cases, we obtain the usual definition of median and quantiles. However, the multivariate case $\mathcal{X} = \mathbb{R}^d$ is more interesting. The generalization of D_1 is called the Tukey depth (or halfspace depth), which is defined as the infimum of the probabilities of all halfspaces containing x. (see Zuo and Serfling, 2000)





Tukey's Depth II

Cuesta-Albertos and Nieto-Reyes (2008) propose a random version of D_T to minimize computational cost. Let us take k projection directions $v_1, \ldots, v_k$ iid.

$$D_{RT}(P, x) = \inf_{1 \leq i \leq k} D_1(P_{v_i}, \langle v_i, x \rangle),$$

where P_{v_i} is the distribution of $\langle v_i, X \rangle$ (the one-dimensional projection of X onto the direction v_i).

This definition can be extended (maintaining almost the same definition) to the case where $\mathcal{X}$ is a separable Hilbert space (for example $L^2[0, 1]$).

This definition can be extended (with minor modifications) to the case where $\mathcal{X}$ is a separable Hilbert space, such as $L^2[0,1]$. In this case, the random vectors v_i are replaced with elements from the continuous dual space $\mathcal{X}^*$. By the Riesz Representation Theorem, these dual elements v are associated with kernel functions a that define a linear continuous operator $x \mapsto \int_0^1 a(s)x(s)ds$, which replaces the finite-dimensional operator $x \mapsto \langle v, x \rangle$. Therefore, choosing the dual elements v corresponds to randomly selecting the corresponding kernel functions a using an appropriate L^2 process.

Spatial depth:

$$SD(x) = 1 - \mathbb{E}\| \frac{x - X}{\|x - X\|} \|$$

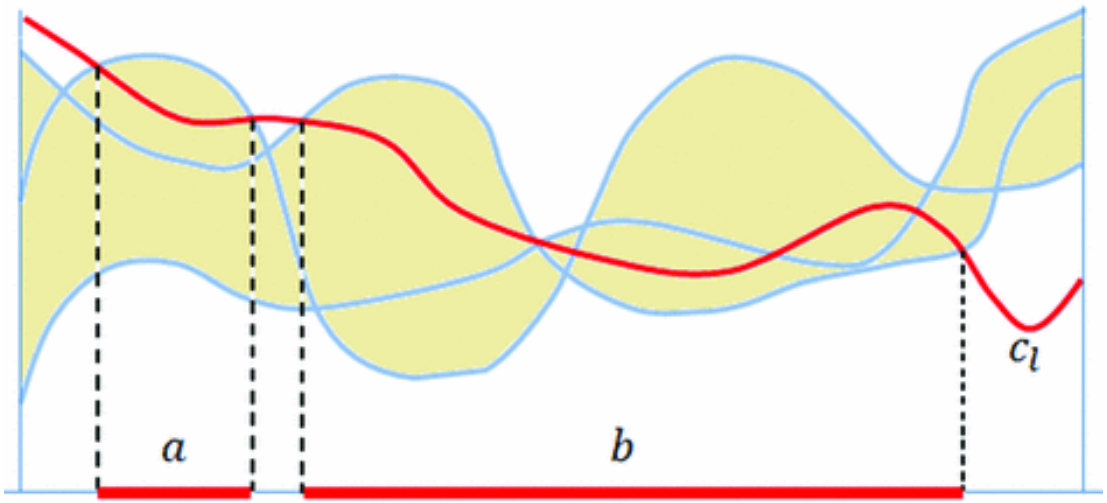
The idea behind the theorem of random projections is that under certain conditions, the probability distribution in a separable Hilbert space is determined by the one-dimensional linear projections onto any set of positive measure. That is, given a Gaussian reference measure μ, if two probability distributions are different, the μ-probability of finding two identically distributed one-dimensional linear projections is zero.

Depth Given a set $X_1, \ldots, X_n$, a random and independent direction is chosen, and the data is projected onto it. Then, the depth of each X_i is defined with some simple one-dimensional depth measure. In the case of functional data, we can assume that we are in $L^2[0,1]$ and use its inner product for the projection. Once this depth estimator is obtained, other directions can be taken and the results averaged.

Depth measures in FDA

An alternative definition is the band depth (BD), introduced by López-Pintado and Romo (2009). To compute this functional depth, one needs to identify the bands that are delimited by all possible pairs of functional observations in the sample. The BD value is the fraction of bands that completely encompass the curve. In *scikit-fda*, this quantity can be computed using methods of the class BandDepth.

A related, less restrictive measure, is the modified band depth (MBD). This measure takes into account not only the number of bands that contain the function, but also the time that $x(t)$ lies within each band. The MBD has better statistical properties than the original BD, in part because it is an integrated depth measure (Nagy et al. 2016). In *scikit-fda*, MBD is implemented in the class ModifiedBandDepth.

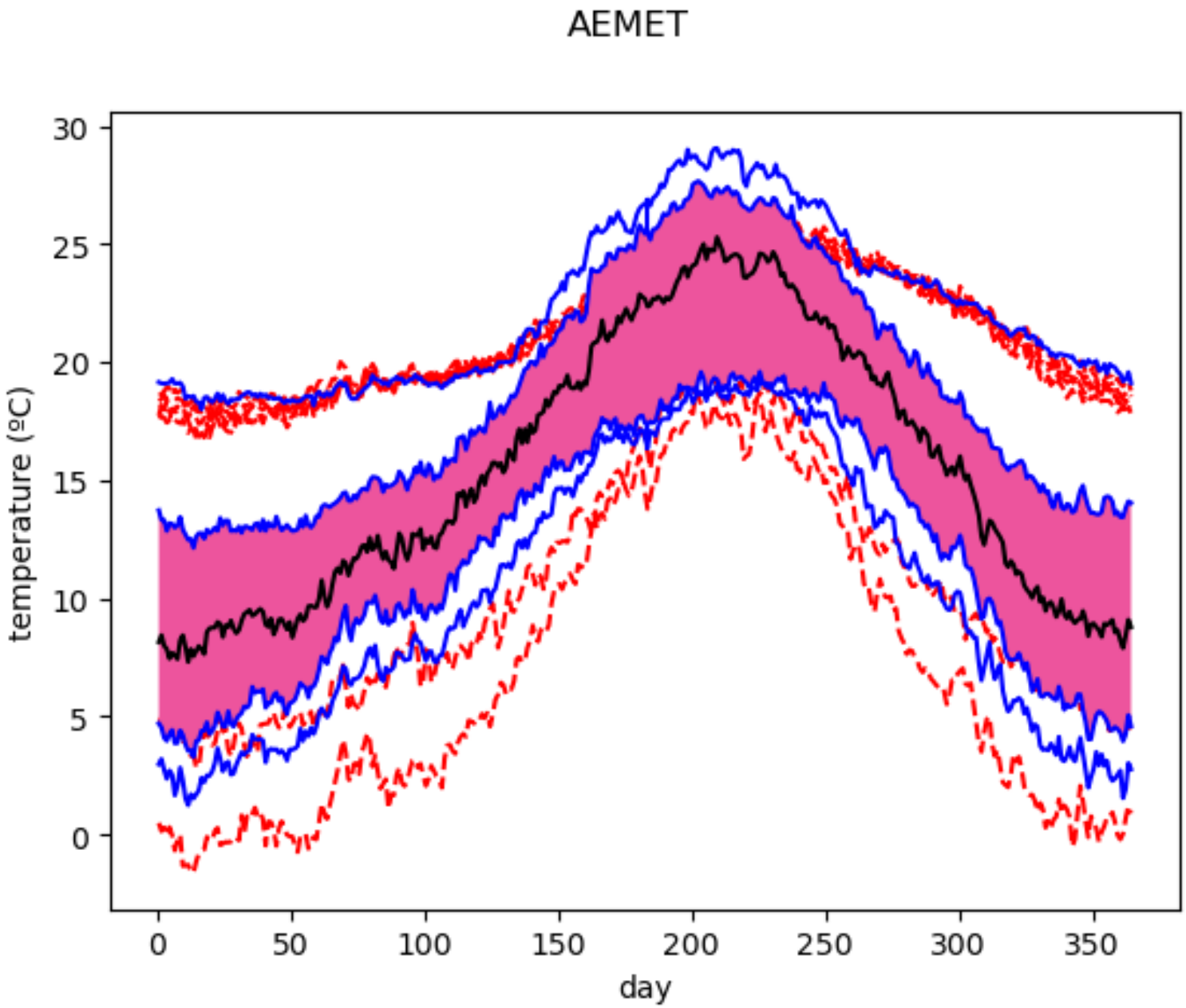


Boxplot funcional

The functional boxplot (Sun and Genton 2011) is a generalization of the univariate boxplot for functional data. I consists of a graph of the functional median surrounded by a central envelope, which encompasses the deepest 50% of the observations, and a maximum non-outlying envelope. The width of this outer envelope is determined by scaling the central one by a constant factor. Its default value is 1.5. but it can be selected by the user. The class Boxplot can be used to generate and customize functional boxplots.

In this plot, a trajectory is marked as an outlier if it lies beyond the maximum non-outlying envelope for some interval. The class BoxplotOutlierDetector can be used for outlier detection based on this criterion. Some customizable elements of Boxplot objects are the depth measure, and the fraction of the deepest observations that define the inner bands.

Boxplot funcional



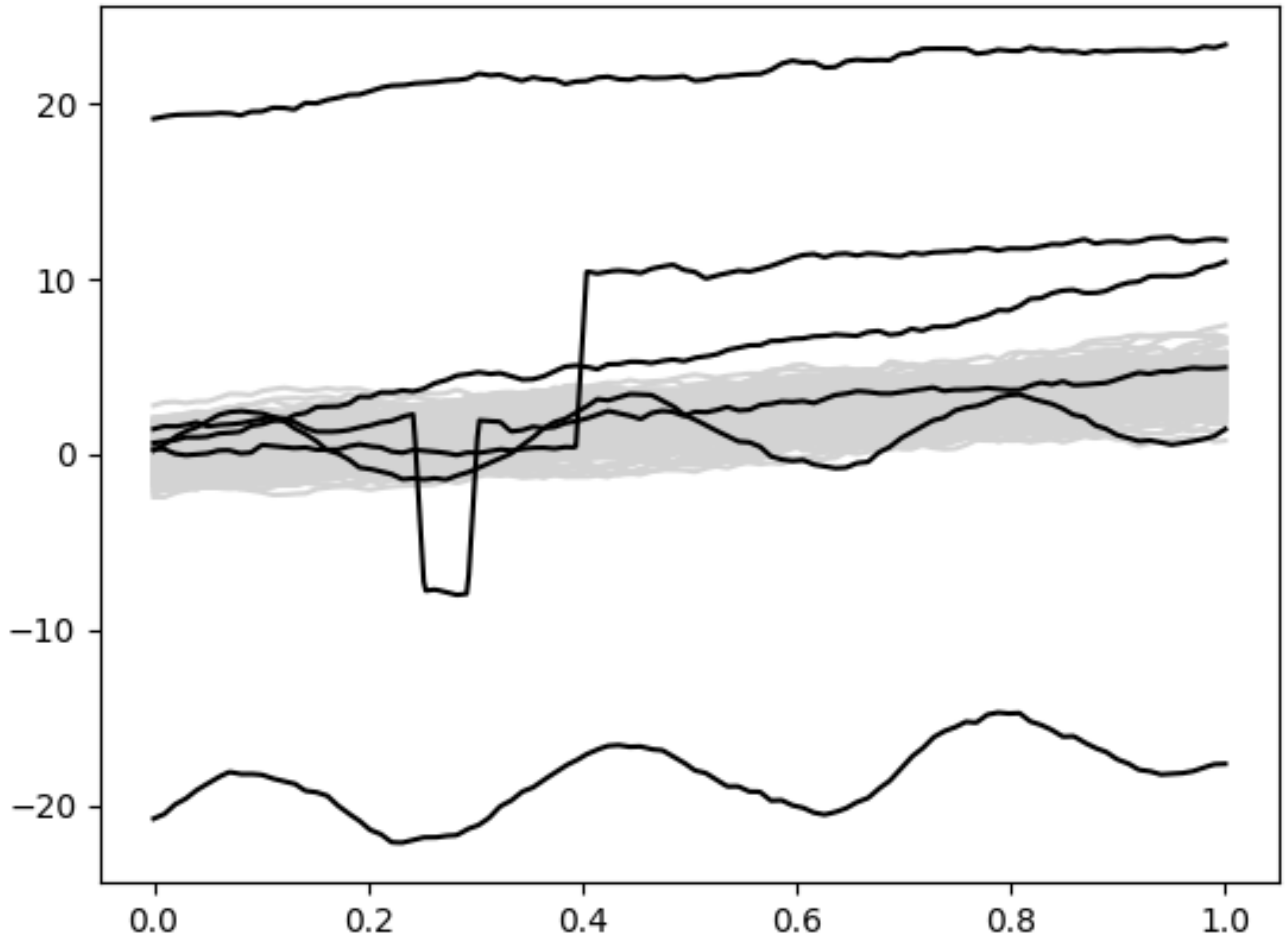
Magnitud-Shape plot

The magnitude-shape (MS-plot) (Dai and Genton 2018, 2019) characterizes the degree of outlyingness of a functional observation is characterized in terms of two quantities: the magnitude outlyingness (MO) and the shape outlyingness (VO).

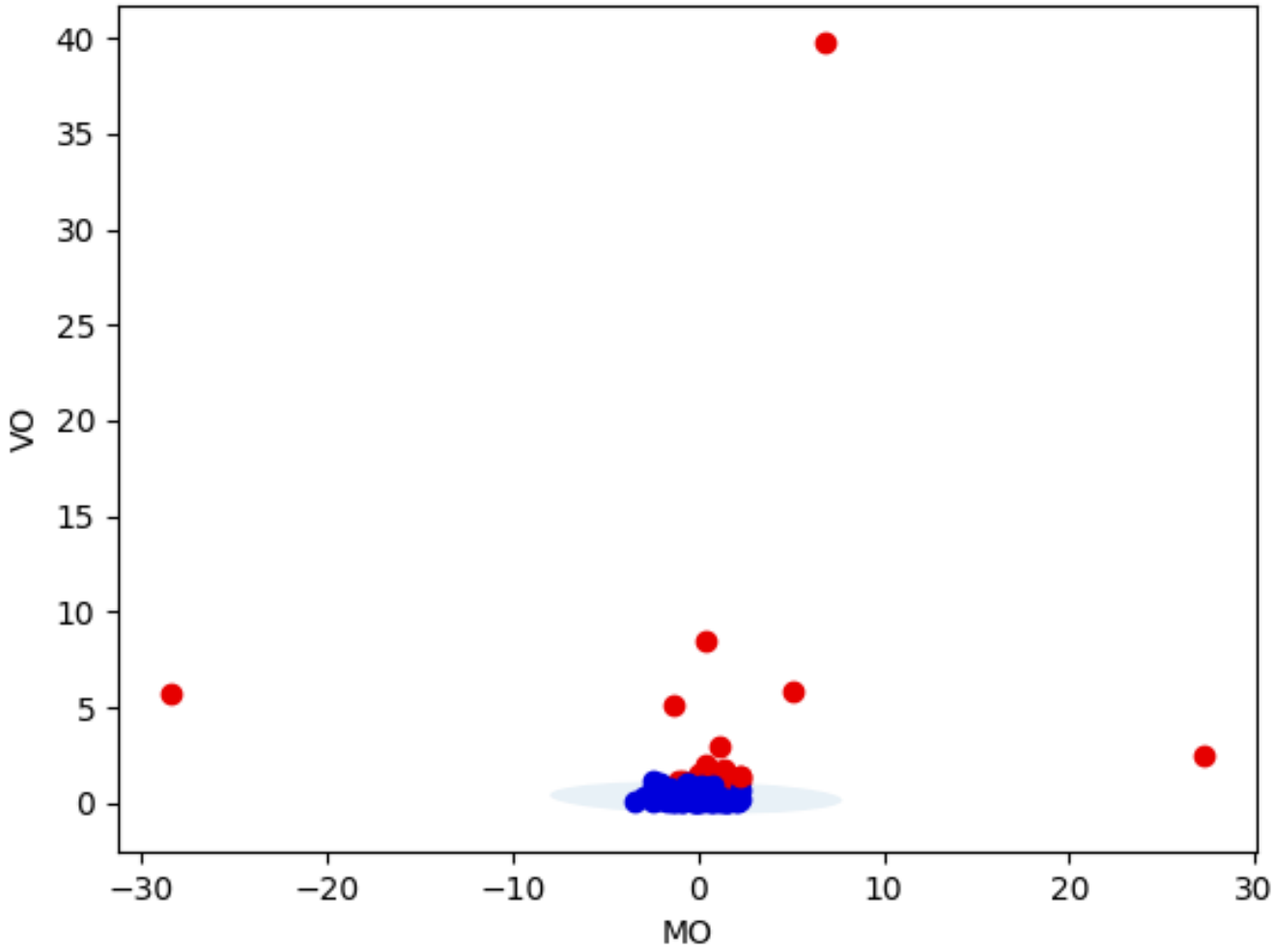
The MS-plot is the scatter plot of the values MO and VO for each functional observation. This two-dimensional representation of the data can be used, for instance, to identify clusters of functions, or detect potential outliers, either in shape or in magnitude.

The class MagnitudeShapePlot generates the MS-plot and uses internally the methods of the class MSPlotOutlierDetector for outlier detection.

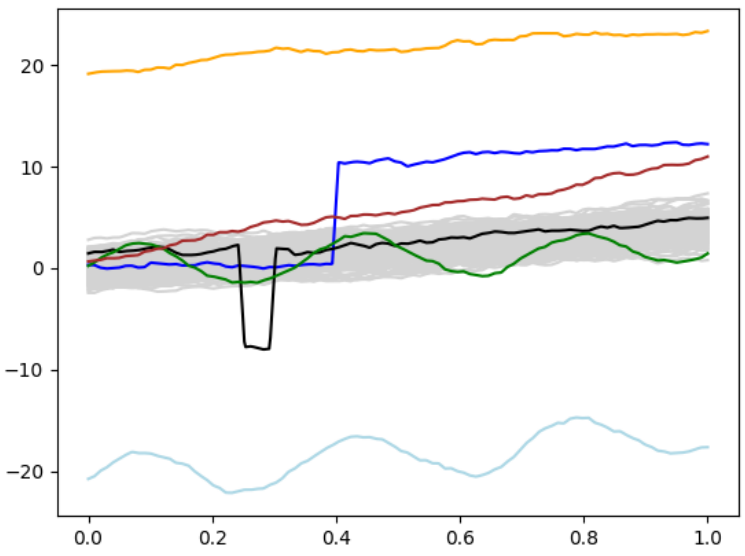
Magnitude-Shape plot

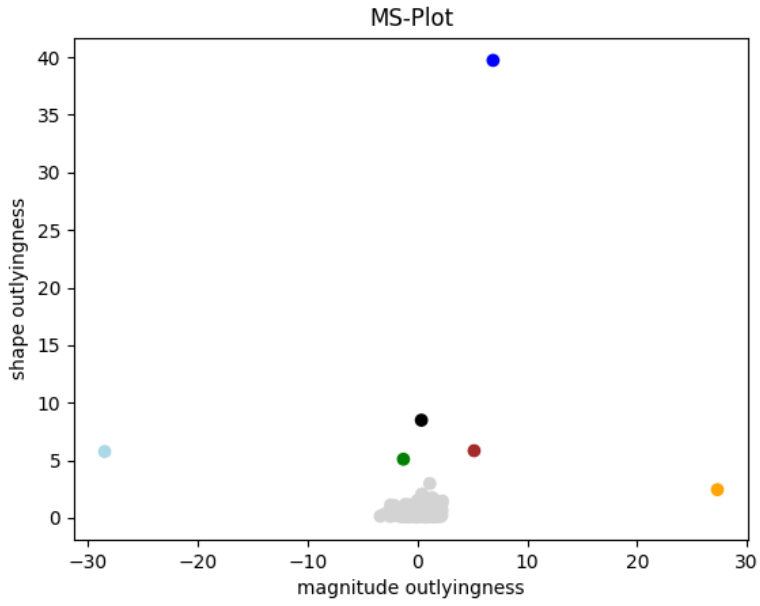


Magnitude-Shape plot



Magnitude-Shape plot





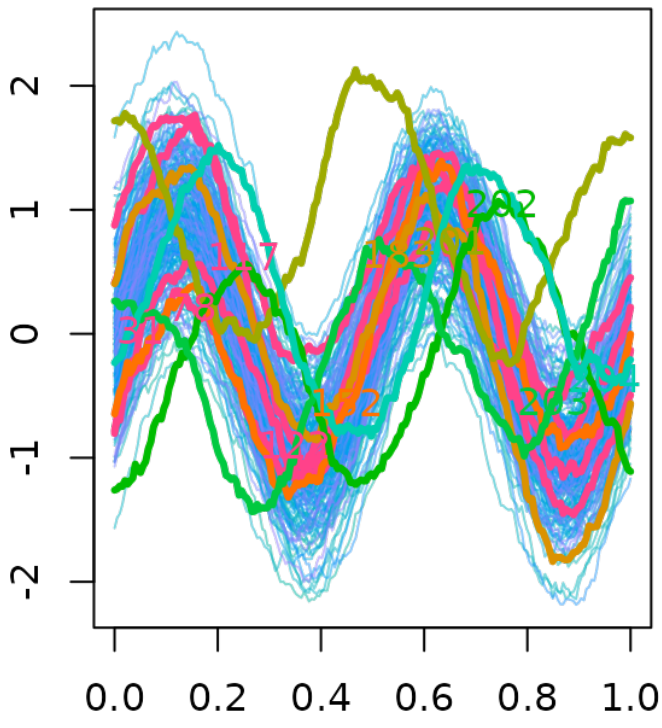
Outliergram

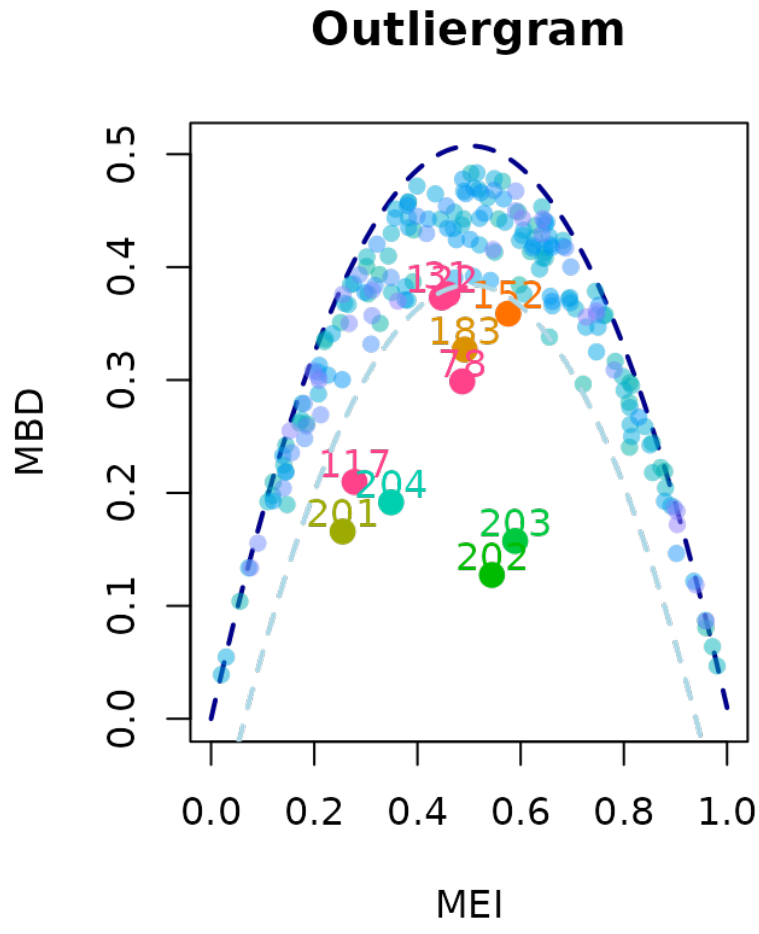
The class Outliergram provides an additional method for data visualization and detection of shape outliers (Arribas-Gil and Romo 2014). The graph is defined in terms of two related quantities: the modified epigraph index (MEI) and the MBD.

The MEI of a trajectory is the average over time of the fraction of curves in the sample that lie above it. Each curve is a point (MEI, MBD) in the scatter plot.

The outliergram takes advantage of the fact that points corresponding to typical functional observations lie on a parabola, whose analytical form is known. This parabola is used as a reference for the identification of shape outliers. Specifically, the degree of outlyingness of a curve is quantified in terms of its vertical distance to the parabola.

Outliergram





Registration

Registration consists in applying transformations to the raw data so that the functional observations are properly aligned.

A number of strategies can be used for registration. For instance, maxima, minima, zeros, and other landmarks can be used as reference. Alternatively, some measure of dispersion between the observations can be minimized. It is also possible to register a set of functional observations to a reference function.

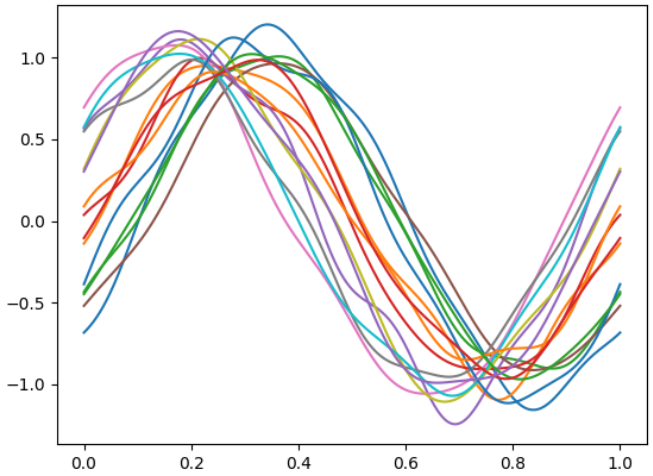
After registration, it may be necessary to evaluate the functional observations at points in the domain that are different from the ones in the original grid. This can be made utilizing the interpolation and extrapolation techniques.

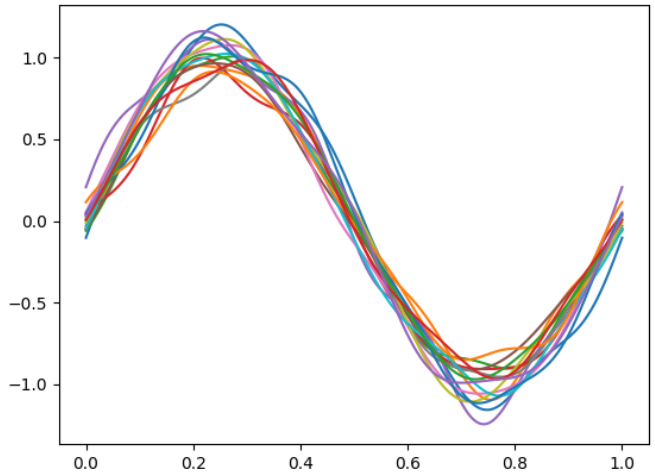
Magnitude-Shape plot

Shift registration consists in aligning the functional observations by a translation

$$\hat{x}_i(t) = x_i(t + \delta_i), \quad i = 1, \ldots, n,$$

The function landmark_shift_registration() can be used to carry out this transformation. The values of the δ_i can be retrieved using the landmark_shift_deltas() function.





Elastic registrationIn elastic

In elastic registration, one attempts to align the data by applying a warping transformation

$$\hat{x}_i(t) = x_i(\gamma_i(t)), \quad i = 1, \ldots, n,$$

In scikit-fda we have landmark_elastic_registration() for elastic landmark registration (the warpings can be retrieved with landmark_elastic_registration_warping()), and FisherRaoElasticRegistration for elastic registration to a common template.