Accepted Manuscript

Local half-region depth for functional data

Claudio Agostinelli

PII: S0047-259X(17)30622-X
DOI: https://doi.org/10.1016/j.jmva.2017.10.004
Reference: YJMVA 4297

To appear in: *Journal of Multivariate Analysis*

Received date : 7 September 2016

Please cite this article as: C. Agostinelli, Local half-region depth for functional data, *Journal of Multivariate Analysis* (2017), https://doi.org/10.1016/j.jmva.2017.10.004

Local half-region depth for functional data

Claudio Agostinelli

Dipartimento di Matematica
Università degli Studi di Trento, Trento, Italy
claudio.agostinelli@unitn.it

Abstract

Data depth has proved successful in the analysis of multivariate data sets, e.g., for deriving an overall center and assigning ranks to the observed units. Two key features are the directions of the ordering, from the center towards the outside, and the recognition of a unique center irrespective of the distribution being unimodal or multimodal. These properties derive from the monotonicity of the ranks that decrease along any ray from the deepest point. Recently, a wider framework allowing for the identification of partial centers was suggested in [2]. The corresponding generalized depth functions, called *local depth functions*, can record local fluctuations and be used for mode detection, identification of components in mixture models, and cluster analysis. As functional data are becoming more common, López-Pintado and Romo [29] recently proposed a notion of half-region depth suited for functional data and for high-dimensional data. Here, we propose a local version of this concept, we study its theoretical properties, we define new similarity measures based on it, and we illustrate its behavior with examples based on real data sets.

Keywords: Clustering, functional data, half-region depth, local depth, similarity measure, time series.

1. Introduction

Functional data, such as continuous trajectories of a process, high frequency time series, and irregularly spaced time series, are encountered in many fields, thanks to the increased sensitivity and recording power of measuring devices and electronic devices. Statistical analysis of functional data calls for specific methods; see [4, 8, 13, 16, 19, 23, 37, 40] for recent overviews. A general task in functional data analysis is to devise an ordering within a sample of curves, with properties similar to (univariate) order statistics. Tools based on data depth often prove useful in this context because they apply to general types of data and are by no means confined to standard multivariate data.

The basic notion in data depth is the concept of depth function, which describes the degree of centrality of the points of the reference space according to the underlying probability distribution. Statistical depth functions are invariant to all (non singular) affine transformations; for symmetric distributions, they reach the maximum value at the center of symmetry and they become negligible when the norm of the point tends to infinity. Another critical property is ray-monotonicity, i.e., statistical depth does not increase along any ray from the center. We refer to [26, 45] for motivation and discussion of these properties.

The classical notion of data depth has been extended to functional data and the various implementations currently available aim to describe the degree of centrality of curves with respect to an underlying probability distribution or a sample. Some desirable properties of functional depths are, e.g., invariance under some class of transformations, vanishing when the norm of a curve tends to infinity, (semi-)continuity, and consistency; see [31] for more details.

Once a depth function has been defined, the main tools for data analysis are depth values and depth regions. Depth values measure the degree of centrality of the points of the space, thus providing a ranking of multivariate observations. This ranking is very different from the usual ranking of the points on the real line, where observations are ordered from the smallest to the largest, because it assigns the highest ranks to the points near the center and the ranks monotonically decrease moving from the center towards the outside. Depth regions enclose the points whose centrality is not less than a fixed value and their location, size and shape convey a lot of information about the underlying distribution. Note that a depth region can be reparameterized so as to depend on the probability content

instead of the depth value. It then follows that the boundary of the depth regions can be interpreted as a quantile surface of the appropriate order [38].

Statistical depth was first introduced for multivariate observations, and many different definitions of depth are currently used; see [27] and the reference therein for a review. One well-known example is Tukey halfspace depth [43], which is the minimum probability of all halfspaces including a point in an Euclidean space

Practical implementations of the data depth paradigm are often computational demanding; such is the case, e.g., for the above depth function [1]. This prevents a straightforward extension of depth functions invented for multivariate data to the context of high-dimensional or functional data. Recently, specific notions of depth for functional data have been introduced which can be adapted to high-dimensional data without a large computational burden. Several contributions to statistical data depths for functional data are present in the literature; see, e.g., Fraiman and Muniz [14] (integrated depth), Cuevas et al. [9] (random projection depth), Cuesta-Albertos and Nieto-Reyes [7] (model depth), [28] (band depth), López-Pintado and Romo [29] (half-region depth), Chakraborty and Chaudhuri [5] (functional spatial depth) and Kuhnt and Rehage [25] (angle-based multivariate functional pseudo-depth). Recent reviews of depth for functional data and their application in classification problems can be found in [6, 30, 34, 35].

In many problems related to functional variables, local features play a major role, mainly because of the infinite-dimensional structure of the data, even more so than in usual multivariate settings. For instance, Demongeot et al. [10] consider a local linear smoothing for the estimation of a regression function; their approach is based on the minimization of the mean square relative error, which is an objective function that highlights the structure of part of the data in contrast with the widely used least square principle where each units have assigned equal weights. Another example is Kara et al. [22], where the kNN regression approach is used to obtain local behavior. However, the notions of depth proposed in the literature so far are not designed to catch local features.

To this aim we propose the concept of local (modified) half-region depth. Goia and Vieu [15] highlight the relation between statistics for functional data and high- (but finite-) dimensional data. The local modified half-region depth falls in this category. We study its properties for infinite- and finite-dimensional data and we illustrate its behavior in both domains. Sguera et al. [39] propose another notion of functional depth which is local-oriented, as well as a kernel-based version of the functional spatial depth. Similarity measures based on local depth are introduced and used in hierarchical clustering procedures.

We provide two examples based on real data sets. The first is the well known Berkeley Growth Study data set [42]. The second is an application in the field of environmental studies, where we consider a wind speed data set obtained from the meteorological station situated in Col De la Roa, San Vito di Cadore, Belluno, Italy (data from this meteorological station are available at `http://intra.tesaf.unipd.it/Sanvito/`). The station records wind speeds regularly every 15 minutes, and we consider the period 2001–2004. After removing 42 incomplete days, we obtain the final wind speed data set composed of 1420 curves which can be interpreted as daily wind speed profiles measured at $4 \times 24 = 96$ time points. The analysis are reported in Section 5. Another application of the local modified half-region depth is available in Agostinelli and Rotondi [3], where the analysis of the shape of macroseismic fields is performed using the concept of local modified half-region depth.

The paper is organized as follows. Section 2 reviews the definitions and main properties of the (modified) half-region depth and Section 3 introduces the local half-region depth and the local modified half-region depth. Section 4 introduces similarity measures based on the new local depths. These measures are used in a hierarchical clustering procedure. Two illustrations involving real data are presented in Section 5 and simulations are reported in Section 6. Section 7 contains concluding remarks. An Online Supplement includes further examples, the properties of the local half-region depth applied to the finite-dimensional case, the proofs of all results presented herein, and an illustration of the R package `ldfun` for the methods presented in this work. The package is available from the author upon request.

2. Background on functional data depth

In this section, we introduce the notation for functional data sets and we discuss the notion of half-region depth. A typical model for functional data is a function $y = y(t)$ with $t \in T$ belonging to the space $C(T)$ of real continuous functions on some compact interval $T \subset \mathbb{R}$. On a probability space $(\Omega, \mathcal{A}, P)$, we consider a stochastic process $Y = \{Y(t) : t \in T\}$ as a family of random variables $Y(t)$ with trajectory in $C(T)$. We let P denote the law of the stochastic process Y. A functional data set is a collection of functions $\mathbf{y}_n = \{y_i \in C(T) : i \in \{1, \ldots, n\}\}$ as a result of n independent trajectories from the stochastic process Y.

The graph of a function y is the subset of the space $G(y) = \{(t, y(t)) : t \in T\}$. The hypograph (hyp) and the epigraph (epi) of a function $y \in C(T)$ are respectively defined as follows:

$$\text{hyp}(y) = \{(t,z) : t \in T, z \leq y(t)\}, \quad \text{epi}(y) = \{(t,z) : t \in T, z \geq y(t)\}.$$

The proportions of graphs that belong to the hypograph and epigraph of a function y are respectively given by

$$R_{hyp}(y; \mathbf{y}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{G(y_i) \subset \text{hyp}(y)\} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\forall_{t\in T}\ y_i(t) \leq y(t)\},$$
$$R_{epi}(y; \mathbf{y}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{G(y_i) \subset \text{epi}(y)\} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\forall_{t\in T}\ y_i(t) \geq y(t)\}.$$

These definitions can be extended in a straightforward way to a stochastic process Y with trajectories in $C(T)$. In this case, we define $R_{hyp}(y; Y) = P\{G(Y) \subset \text{hyp}(y)\}$ and $R_{epi}(y; Y) = P\{G(Y) \subset \text{epi}(y)\}$. A simple notion of functional depth is the minimum probability that a trajectory of the random process Y belongs to the hypograph or to the epigraph of a given curve y.

Definition 1 (Half-region depth, [29])**.** *For a functional data set $\mathbf{y}_n$, the half-region depth of a curve y is $d_{HR}(y; \mathbf{y}_n) = \min\{R_{hyp}(y; \mathbf{y}_n), R_{epi}(y; \mathbf{y}_n)\}$. The population version is $d_{HR}(y; Y) = \min\{R_{hyp}(y; Y), R_{epi}(y; Y)\}$.*

Given a probability space $(\Omega, \mathcal{A}, P)$, where Ω is a sample space, $\mathcal{A}$ is a σ-algebra on Ω and P a probability measure on the measurable space $(\Omega, \mathcal{A})$, we consider a random vector $\mathbf{X} \in \mathbb{R}^p$ with law $\Pr_{\mathbf{X}}$ and p a strictly positive integer.

The notions of hypograph and epigraph can be adapted to finite-dimensional data. Parallel coordinates [20, 32] are a convenient tool to visualize a set of points in $\mathbb{R}^p$. The Cartesian orthogonal axes become parallel and equally spaced in parallel coordinates; thus, points with dimension larger than 3 can be easily represented. Observations in $\mathbb{R}^p$ can be seen as real functions defined on the set of indices $1, \ldots, p$ and expressed as $\mathbf{x} = (x(1), \ldots, x(p))$. The hypograph and epigraph of a point $\mathbf{x} \in \mathbb{R}^p$ can be expressed respectively as

$$\text{hyp}(\mathbf{x}) = \{(k,z) \in (1,\ldots,p) \times \mathbb{R} : z \leq x(k)\}, \qquad \text{epi}(\mathbf{x}) = \{(k,z) \in (1,\ldots,p) \times \mathbb{R} : z \geq x(k)\}.$$

In a similar way, we can define the graph $G(\mathbf{x}) = \{(k, x(k)) \in \{1,\ldots,p\} \times \mathbb{R}\}$. For each $j \in \{1,\ldots,p\}$, let $\mathbf{e}_j$ be a $p \times 1$ vector with 1 at the jth coordinate and 0 otherwise. The half-region depth in Cartesian coordinates on $\mathbb{R}^p$ for the random vector $\mathbf{X}$ at the point $\mathbf{x}$ is given by

$$d_{HR}(\mathbf{x}; \mathbf{X}) = \min\left[\Pr_{\mathbf{X}}\{G(\mathbf{X}) \in \text{hyp}(\mathbf{x})\}, \Pr_{\mathbf{X}}\{G(\mathbf{X}) \in \text{epi}(\mathbf{x})\}\right] = \min\left[\Pr_{\mathbf{X}}\left[\bigcap_{j=1}^{p} \text{HS}_{e_j}\{x(j)\}\right], \Pr_{\mathbf{X}}\left[\bigcap_{j=1}^{p} \text{HS}_{-e_j}\{x(j)\}\right]\right].$$

The properties of the concept of half-region depth are studied in [29]. In the finite-dimensional case, half-region depth is not affine invariant, though it is invariant with respect to translations and some types of dilation, i.e., linear transformations operated by positive (or negative) definite diagonal matrices. For $p = 1$, half-region depth is equivalent to the halfspace depth. It vanishes as the norm of $\mathbf{x}$ increases. The empirical version uniformly almost surely converges to the corresponding population version, the empirical maximizer is a consistent estimator of the unique maximizer of the population version. In the general case half-region depth is invariant under the transformations $aY + b$, where $a(t)$ is either positive or negative in T. Similar results, as for the finite-dimensional case, hold for the asymptotic behavior.

A modified version of half-region depth which is less restrictive than the definition above is proposed in [29]. This modified version can be used for the analysis of a set of curves with many crossing points; this is frequent when the curves are irregular (non-smooth). We denote the superior (EL) and the inferior (HL) lengths respectively as

$$\text{EL}(y; Y) = \frac{1}{\lambda(T)}\text{E}[\lambda\{t \in T : y(t) \leq Y(t)\}], \quad \text{HL}(y; Y) = \frac{1}{\lambda(T)}\text{E}[\lambda\{t \in T : y(t) \geq Y(t)\}],$$

where λ stands for the Lebesgue measure on $\mathbb{R}$. $\text{EL}(y; Y)$ can be interpreted as the "proportion of time" the stochastic process Y is greater than the curve y, similarly for $\text{HL}(y; Y)$. The modified half-region depth at y is

$$d_{MHR}(y; Y) = \min\{\text{EL}(y; Y), \text{HL}(y; Y)\}.$$

Let $\mathbf{y}_n = (y_1, \ldots, y_n)$ be a set of curves from the stochastic process Y. The sample version of this depth is

$$d_{MHR}(y; \mathbf{y}_n) = \min\{\mathrm{EL}(y; \mathbf{y}_n), \mathrm{HL}_n(y; \mathbf{y}_n)\},$$

where

$$\mathrm{EL}(y; \mathbf{y}_n) = \frac{1}{n\lambda(T)} \sum_{i=1}^{n} \lambda\{t \in T : y(t) \leq y_i(t)\}, \quad \mathrm{HL}(y; \mathbf{y}_n) = \frac{1}{n\lambda(T)} \sum_{i=1}^{n} \lambda\{t \in T : y(t) \geq y_i(t)\}.$$

are the sample means of the proportions of the lengths in the epigraph and hypograph, respectively.

3. Local functional depths

Local depths [2] are a recent proposal to extend the classical definition of statistical data depth. In early days, it was often assumed that depth ranks could single out just one center of a distribution, corresponding to the maximizer of the ranks, whether the distribution is unimodal or multimodal. Current developments are showing that local depths can indeed account for multimodal data, having multiple centers. These generalized definitions measure centrality conditional on a neighborhood of each point of the space and provide a tool that is sensitive to local features of the data, while retaining most features of regular depth functions. In most cases, when the neighborhood radius tends to infinity, the usual (global) depth is recovered. However, when the distribution is unimodal, local depth behaves very similarly to a regular depth and provides a similar ranking. Local halfspace depth is introduced in [2]; halfspaces are replaced by infinite slabs with finite width. When the threshold (slab width) tends to infinity, the ordinary definition is recovered.

Hereafter, we introduce and study the properties of the local half-region depth and the local modified half-region depth for functional data. The properties of the local half-region in finite-dimensional space are also studied.

3.1. Local half-region depth

A local version of half-region depth is obtained by imposing a restriction on the size of the hypograph and epigraph under consideration. For a given non-negative function $\tau = \tau(t)$ with $t \in T$, the closed negative slab $\mathrm{hyp}(y; \tau)$ is the intersection between $\mathrm{hyp}(y)$ and $\mathrm{epi}(y - \tau)$, i.e.,

$$\mathrm{hyp}(y; \tau) = \mathrm{hyp}(y) \cap \mathrm{epi}(y - \tau) = \{(t, z) : t \in T, y(t) - \tau(t) \leq z \leq y(t)\}$$

and, similarly for the closed positive slab $\mathrm{epi}(y; \tau)$. The local half-region depth is defined as the minimum probability of these two sets.

Definition 2 (Local half-region depth)**.** *For a stochastic process Y, the local half-region depth for the curve y is defined as $ld_{HR}(y; Y, \tau) = \min[P\{G(Y) \subset \mathrm{hyp}(y; \tau)\}, P\{G(Y) \subset \mathrm{epi}(y; \tau)\}]$. For a functional data set $\mathbf{y}_n$, the empirical version is*

$$ld_{HR}(y; \mathbf{y}_n, \tau) = \min\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{G(y_i) \subset \mathrm{hyp}(y; \tau)\}, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{G(y_i) \subset \mathrm{epi}(y; \tau)\}\right].$$

In the finite-dimensional case and Cartesian coordinates, this leads to consider slabs instead of halfspaces, each slab $j \in \{1, \ldots, p\}$ has its own dimension given by $\tau(j)$, viz.

$$ld_{HR}(\mathbf{x}; \mathbf{X}, \boldsymbol{\tau}) = \min\left[\Pr_{\mathbf{X}}\left\{\bigcap_{j=1}^{p} \mathrm{SL}_{-e_j}(x(j) - \tau(j), x(j))\right\}, \Pr_{\mathbf{X}}\left\{\bigcap_{j=1}^{p} \mathrm{SL}_{e_j}(x(j), x(j) + \tau(j))\right\}\right].$$

Given a function $F : \mathbb{R}^p \to [0, 1]$ and a point $\mathbf{a} = (a_1, \ldots, a_p) \in \mathbb{R}^p$, consider the restriction $\tilde{F}$ of F to the set $A = \{\mathbf{x} \in \mathbb{R}^p : \forall_{j \in \{1,\ldots,p\}}\ x_j < a_j\}$. We denote by $F(\mathbf{a}^-)$ the limit of the restriction $\tilde{F}$ at $\mathbf{a}$, i.e., $\lim_{\|\mathbf{x}-\mathbf{a}\| \downarrow 0} \tilde{F}(\mathbf{x})$. A

different formulation is possible using an inclusion-exclusion formula. In fact, the probability of the hypercube can be rewritten in terms of the distribution function $F_{\mathbf{X}}$ of the random vector $\mathbf{X}$ as follows:

$$\Pr_{\mathbf{X}}\left[\bigcap_{j=1}^{p} \mathrm{SL}_{-e_j}\{x(j)-\tau(j), x(j)\}\right] = F_{\mathbf{X}}(\mathbf{x}) + \sum_{k=1}^{p}(-1)^k \sum_{I\subseteq\{1,\ldots,p\},\,|I|=k} F_{\mathbf{X}}\{(\mathbf{x}-\boldsymbol{\tau}_I)^-\},$$

$$\Pr_{\mathbf{X}}\left[\bigcap_{j=1}^{p} \mathrm{SL}_{e_j}\{x(j), x(j)+\tau(j)\}\right] = F_{\mathbf{X}}\{\mathbf{x}+\tau(j)\} + \sum_{k=1}^{p}(-1)^k \sum_{I\subseteq\{1,\ldots,p\},\,|I|=k} F_{\mathbf{X}}\{(\mathbf{x}+\boldsymbol{\tau}_I)^-\},$$

where $\boldsymbol{\tau}_I$ is a $p \times 1$ vector with the elements indexed by I equal to the corresponding element in $\boldsymbol{\tau}$ and 0 for all the others. Furthermore, $|\cdot|$ denotes the cardinality of the set.

Properties of the local half-region depth applied to functional data are now studied; Section SM–1, available in the Online Supplement, provides results for the finite-dimensional case. Let $\mathbf{y}_n = (y_1, \ldots, y_n)$ a functional data set from a stochastic process Y in $C(T)$ with law P. Assume that the stochastic process Y is tight, i.e., $P(\|Y\|_\infty \geq M) \to 0$ as $M \to \infty$.

The local half-region depth satisfies a linear invariance property provided that the threshold function τ is transformed appropriately, i.e., consider a and b functions in $C(T)$, where $a(t) > 0$ or $a(t) < 0$ for every $t \in T$, then $ld_{HR}(ay + b; aY + b, |a|\tau) = ld_{HR}(y; Y, \tau)$. In the next proposition we study the behavior of the local half-region depth according to the behavior of the function τ.

Proposition 3. *Let $ld_{HR}(\cdot, Y, \tau)$ be the local half-region depth. Then, for any given function $y \in C(T)$ and any non-negative functions $\tau_1, \tau_2, \tau \in C(T)$, the following results hold true:*

(i) if $\tau_1 < \tau_2$, then $ld_{HR}(y; Y, \tau_1) \leq ld_{HR}(y; Y, \tau_2)$;

(ii) $P\{G(Y) = g(y)\} \leq ld_{HR}(y; Y, \tau) \leq d_{HR}(y; Y)$;

(iii) $\lim_{\min_{t\in T}\tau(t)\to\infty} ld_{HR}(y; Y, \tau) = d_{HR}(y; Y)$;

(iv) $\lim_{\|\tau\|_\infty\to 0^+} ld_{HR}(y; Y, \tau) = P\{G(Y) = g(y)\}$.

Proofs are reported in Section SM–7 of the Online Supplement. In the analysis of a functional data set, the function τ should be determined. In most cases τ can be the constant function equal to a quantile of the empirical distribution of the distance between all pairs of curves. In practice, the use of the sup norm to measure the distance and a quantile order in the interval 5%–30% proved to be effective in most situations.

The local half-region depth of a function converges to zero when its norm tends to infinity.

Proposition 4. *The local half-region depths $ld_{HR}(y; Y, \tau)$ and $ld_{HR}(y; \mathbf{y}_n, \tau)$, for a non-negative function $\tau \in C(T)$, satisfy*

$$\lim_{M\to\infty} \sup_{\|y\|_\infty \geq M} ld_{HR}(y; Y, \tau) = 0, \quad \lim_{M\to\infty} \sup_{\|y\|_\infty \geq M} ld_{HR}(y; \mathbf{y}_n, \tau) = 0.$$

The local half-region depth is a continuous function when the marginal distributions of the random process Y are absolutely continuous. We recall that a function $y(t)$ of the real line is upper semicontinuous at t if, for each $\epsilon > 0$, there exists $\delta > 0$ such that $|t - s| < \delta$ implies that $y(s) < y(t) + \epsilon$.

Proposition 5. *Let $\tau \in C(T)$ be a non-negative function. Then $ld_{HR}(\cdot; Y, \tau)$ is an upper semicontinuous functional. Moreover, if P has absolutely continuous marginals, then $ld_{HR}(\cdot; Y, \tau)$ is continuous.*

In the next theorem, we establish the strong consistency of the sample local half-region depth.

Theorem 6. *For a non-negative function $\tau \in C(T)$, $ld_{HR}(\cdot; \mathbf{Y}_n, \tau)$ is strongly consistent: $ld_{HR}(y; \mathbf{Y}_n, \tau) \xrightarrow{a.s.} ld_{HR}(y; Y, \tau)$.*

Hereafter, we establish the uniform consistency of the sample local half-region depth and the strong consistency of its global maximizer using the approach based on empirical processes; see, e.g., [24]. We recall that a subset E of $C(T)$ is equicontinuous if, for each $\nu > 0$, there exists $\delta(\nu) > 0$ such that, for every $y \in E$ and for every $t, s \in T$, if $|t - s| < \delta(\nu)$, then $|y(t) - y(s)| < \nu$.

Theorem 7. *Let E be an equicontinuous subset of $C(T)$ and assume that the stochastic process Y satisfies the following condition: (i) for a given $\varepsilon > 0$, there exists $\nu(\varepsilon) > 0$ such that for every pair of functions $z_i, z_j \in C(T)$, if $\|z_i - z_j\|_\infty \leq \nu(\varepsilon)$, then $P(z_j \leq Y \leq z_i) \leq \varepsilon$; and (ii) given a non-negative function $\tau \in C(T)$ such that $\inf_{t\in T} \tau(t) \geq \underline{\tau} > 2\nu(\varepsilon) > 0$ for a constant $\underline{\tau}$, then $ld_{HR}(\cdot; \mathbf{Y}_n, \tau)$ is strongly uniform consistent, i.e.,*

$$\sup_{y\in E} |ld_{HR}(y; \mathbf{Y}_n, \tau) - ld_{HR}(y; Y, \tau)| \xrightarrow{a.s.} 0.$$

The following theorem states the convergence of the global maximizer of the sample local half-region depth.

Theorem 8. *Under the assumptions stated in the previous theorem, if $ld_{HR}(y; Y, \tau)$ is uniquely maximized at $\hat{y} \in E$, i.e., $\hat{y} = \arg\max_{y\in E} ld_{HR}(y, Y, \tau)$ and $\hat{Y}_n$ is a sequence of functions in E such that $\hat{Y}_n = \arg\max_{y\in E} ld_{HR}(y, \mathbf{Y}_n, \tau)$, then $\hat{Y}_n \xrightarrow{a.s.} \hat{y}$ when $n \to \infty$.*

3.2. Local modified half-region depth

To introduce a modified version of local half-region depth, the expected proportions of time a process (or a function) stays in the upper and lower slabs need be defined:

$$\mathrm{EL}(y; Y, \tau) = \frac{1}{\lambda(T)} \mathrm{E}\left[\{\lambda(t \in T : y(t) \leq Y(t) \leq y(t) + \tau(t))\} \mathbf{1}\{\forall_{t\in T}\ y(t) - \tau(t) \leq Y(t) \leq y(t) + \tau(t)\}\right],$$
$$\mathrm{HL}(y; Y, \tau) = \frac{1}{\lambda(T)} \mathrm{E}\left[\{\lambda(t \in T : y(t) - \tau(t) \leq Y(t) \leq y(t))\} \mathbf{1}\{\forall_{t\in T}\ y(t) - \tau(t) \leq Y(t) \leq y(t) + \tau(t)\}\right],$$

where $\mathbf{1}$ denotes an indicator function. The corresponding quantities for a functional data set $\mathbf{y}_n$ are

$$\mathrm{EL}(y; \mathbf{y}_n, \tau) = \frac{1}{n\lambda(T)} \sum_{i=1}^{n} \lambda\{t \in T : y(t) \leq y_i(t) \leq y(t) + \tau(t)\} \mathbf{1}\{\forall_{t\in T}\ y(t) - \tau(t) \leq y_i(t) \leq y(t) + \tau(t)\},$$
$$\mathrm{HL}(y; \mathbf{y}_n, \tau) = \frac{1}{n\lambda(T)} \sum_{i=1}^{n} \lambda\{t \in T : y(t) - \tau(t) \leq y_i(t) \leq y(t)\} \mathbf{1}\{\forall_{t\in T}\ y(t) - \tau(t) \leq y_i(t) \leq y(t) + \tau(t)\}.$$

The local modified half-region depth is the minimum expected length between stay above or below a certain curve.

Definition 9 (Local modified half-region depth)**.** *For a stochastic process Y, the local modified half-region depth for the curve y is defined as $ld_{MHR}(y; Y, \tau) = \min\{\mathrm{EL}(y; Y, \tau), \mathrm{HL}(y; Y, \tau)\}$, and for a functional data set $\mathbf{y}_n$, the empirical version is $ld_{MHR}(y; \mathbf{y}_n, \tau) = \min\{\mathrm{EL}(y; \mathbf{y}_n, \tau), \mathrm{HL}_n(y; \mathbf{y}_n, \tau)\}$.*

Properties of the local modified half-region depth applied to functional data are now studied. The local modified half-region depth satisfies the same linear invariance property of the local half-region depth, i.e., consider a and b functions in $C(T)$, where $a(t) > 0$ or $a(t) < 0$ for every $t \in T$, then $ld_{MHR}(ay + b; aY + b, |a|\tau) = ld_{MHR}(y; Y, \tau)$. In the next proposition we study the behavior of the local modified half-region depth according to the behavior of the function τ.

Proposition 10. *Let $ld_{MHR}(\cdot, Y, \tau)$ be the local modified half-region depth. Then, for any given function $y \in C(T)$ and any non-negative functions $\tau_1, \tau_2, \tau \in C(T)$, the following results hold true:*

(i) if $\tau_1 < \tau_2$, then $ld_{MHR}(y; Y, \tau_1) \leq ld_{MHR}(y; Y, \tau_2)$;

(ii) $P\{G(Y) = g(y)\} \leq ld_{MHR}(y; Y, \tau) \leq d_{MHR}(y; Y)$;

(iii) $\lim_{\min_{t\in T} \tau(t)\to\infty} ld_{MHR}(y; Y, \tau) = d_{MHR}(y; Y)$;

(iv) $\lim_{\|\tau\|_\infty \to 0^+} ld_{MHR}(y; Y, \tau) = P\{G(Y) = g(y)\}$.

Proofs are reported in Section SM–7 of the Online Supplement. The local modified half-region depth of a function converges to zero when its norm tends to infinity.

Proposition 11. *The local modified half-region depths $ld_{MHR}(y; Y, \tau)$ and $ld_{MHR}(y; \mathbf{y}_n, \tau)$, for a non-negative function $\tau \in C(T)$, satisfy*

$$\lim_{M\to\infty} \sup_{\|y\|_\infty \geq M} ld_{MHR}(y; Y, \tau) = 0, \quad \lim_{M\to\infty} \sup_{\|y\|_\infty \geq M} ld_{MHR}(y; \mathbf{y}_n, \tau) = 0.$$

The local modified half-region depth is a continuous function when the marginal distributions of the random process Y are absolutely continuous.

Proposition 12. *Let $\tau \in C(T)$ be a non-negative function. Then $ld_{MHR}(\cdot; Y, \tau)$ is an upper semicontinuous functional. Moreover, if P has absolutely continuous marginals, then $ld_{MHR}(\cdot; Y, \tau)$ is continuous.*

In the next theorem, we establish the strong consistency of the sample local modified half-region depth.

Theorem 13. *For a non-negative function $\tau \in C(T)$, $ld_{MHR}(\cdot; \mathbf{Y}_n, \tau)$ is strongly consistent: $ld_{MHR}(y; \mathbf{Y}_n, \tau) \xrightarrow{a.s.} ld_{MHR}(y; Y, \tau)$.*

Next, we establish the uniform consistency of the sample local modified half-region depth and the strong consistency of its global maximizer.

Theorem 14. *Let E be an equicontinuous subset of $C(T)$ and assume that the stochastic process Y satisfies the following condition: (i) for a given $\varepsilon > 0$, there exists $\nu(\varepsilon) > 0$ such that for every pair of functions $z_i, z_j \in C(T)$, if $\|z_i - z_j\|_\infty \leq \nu(\varepsilon)$, then $P(z_j \leq Y \leq z_i) \leq \varepsilon$; and (ii) given a non-negative function $\tau \in C(T)$ such that $\inf_{t\in T} \tau(t) \geq \underline{\tau} > 2\nu(\varepsilon) > 0$ for a constant $\underline{\tau}$, then $ld_{MHR}(\cdot; \mathbf{Y}_n, \tau)$ is strongly uniform consistent, i.e.,*

$$\sup_{y\in E} |ld_{MHR}(y; \mathbf{Y}_n, \tau) - ld_{MHR}(y; Y, \tau)| \xrightarrow{a.s.} 0.$$

The proof is very similar to that of Theorem 6 and is not reported. The following theorem gives the convergence of the global maximizer of the sample local modified half-region depth.

Theorem 15. *Under the assumptions stated in the previous theorem, if $ld_{MHR}(y; Y, \tau)$ is uniquely maximized at $\hat{y} \in E$, i.e., $\hat{y} = \arg\max_{y\in E} ld_{MHR}(y, Y, \tau)$ and $\hat{Y}_n$ is a sequence of functions in E such that $\hat{Y}_n = \arg\max_{y\in E} ld_{MHR}(y, \mathbf{Y}_n, \tau)$, then $\hat{Y}_n \xrightarrow{a.s.} \hat{y}$ when $n \to \infty$.*

The proof is essentially the same of that of Theorem 8 and is not reported.

4. Hierarchical clustering based on data depth

While the classification problem (supervised learning) is an area of active research in statistical data depth — see, e.g., [6, 30] for update reviews and contributions — the clustering problem (unsupervised learning) has received much less attention. The few contributions in the literature involve the depth ranks in a k-mean algorithm; examples are [11, 21, 41]. Hierarchical clustering [12, 18] plays a very important role in exploratory data analysis and one of its mean features is the nonparametric setting. One necessary ingredient is a dissimilarity/distance measure matrix among the observations. Hereafter we are going to first introduce similarity measures based on introduced local depths; we will then discuss one possible way of transforming this similarity measure into a dissimilarity measure. Given to curves x and y, for the half-region similarity depth we check if both curves belong to a hypograph or an epigraph.

Definition 16 (Half-region similarity depth)**.** *The half-region similarity depth $s_{HR}(x, y; \mathbf{y}_n)$ of the trajectories pair (x, y), given a functional data set $\mathbf{y}_n$, is defined as*

$$s_{HR}(x, y; \mathbf{y}_n) = \min\{R_{hyp}(x \wedge y; \mathbf{y}_n), R_{epi}(x \vee y; \mathbf{y}_n)\},$$

where $x \wedge y = \{(t, z(t)) : t \in T, z(t) = \min\{x(t), y(t)\}\}$, $x \vee y = \{(t, z(t)) : t \in T, z(t) = \max\{x(t), y(t)\}\}$. The population version is given by $s_{HR}(x, y; Y) = \min\{R_{hyp}(x \wedge y; Y), R_{epi}(x \vee y; Y)\}$.

To obtain a local half-region similarity depth $ls_{HR}(x, y; \mathbf{y}_n)$ of the trajectories pair (x, y) given a functional data set $\mathbf{y}_n$, we consider a closed negative slab and a closed positive slab using $x \wedge y$ and $x \vee y$ so that only curves which are not too far to these two curves are considered. More formally we have the following definition.

Definition 17 (Local half-region similarity depth)**.** *The local half-region similarity depth $ls_{HR}(x, y; \mathbf{y}_n)$ of the trajectories pair (x, y), given a functional data set $\mathbf{y}_n$, is defined as*

$$ls_{HR}(x, y; \mathbf{y}_n, \tau) = \min\{R_{hyp}(x \wedge y; \mathbf{y}_n, \tau), R_{epi}(x \vee y; \mathbf{y}_n, \tau)\},$$

and the population version is given by $ls_{HR}(x, y; Y, \tau) = \min\{R_{hyp}(x \wedge y; Y, \tau), R_{epi}(x \vee y; Y, \tau)\}$.

Note that, $ls_{HR}(y, y; \mathbf{y}_n, \tau) = ld_{HR}(y; \mathbf{y}_n, \tau)$ and $ls_{HR}(x, y; \mathbf{y}_n, \tau) \leq \min\{ld_{HR}(x; \mathbf{y}_n, \tau), ld_{HR}(y; \mathbf{y}_n, \tau)\}$ for all $\tau > 0$. The same holds for the population versions. Let $z = x \vee y$ and $w = x \wedge y$. Applying the same idea as for the half-region similarity depth, we can obtain similarity measure based on modified half-region similarity depth.

Definition 18 (Modified half-region similarity depth)**.** *The sample version of the modified half-region similarity depth for the couple (x, y) of trajectories, given the functional data set $\mathbf{y}_n$, is defined as*

$$s_{MHR}(x, y; \mathbf{y}_n) = \min\{\mathrm{EL}(z, w; \mathbf{y}_n), \mathrm{HL}(z, w; \mathbf{y}_n)\},$$

where

$$\mathrm{EL}(z, w; \mathbf{y}_n) = \frac{1}{n\lambda(T)} \sum_{i=1}^{n} \lambda\{t \in T : z(t) \leq y_i(t)\}, \quad \mathrm{HL}(z, w; \mathbf{y}_n) = \frac{1}{n\lambda(T)} \sum_{i=1}^{n} \lambda\{t \in T : y_i(t) \leq w(t)\}$$

and the population version is given by $s_{MHR}(x, y; Y) = \min\{\mathrm{EL}(z, w; Y), \mathrm{HL}(z, w; Y)\}$, where

$$\mathrm{EL}(z, w; Y) = \frac{1}{\lambda(T)} \mathrm{E}[\lambda\{t \in T : z(t) \leq Y(t)\}], \quad \mathrm{HL}(z, w; Y) = \frac{1}{\lambda(T)} \mathrm{E}[\lambda\{t \in T : Y(t) \leq w(t)\}].$$

In the same way, we can have a local modified half-region similarity depth.

Definition 19 (Local modified half-region similarity depth)**.** *The sample version of the local modified half-region similarity depth for the couple (x, y) of trajectories, given the functional data set $\mathbf{y}_n$, is defined as*

$$ls_{MHR}(x, y; \mathbf{y}_n, \tau) = \min\{\mathrm{EL}(z, w; \mathbf{y}_n, \tau), \mathrm{HL}(z, w; \mathbf{y}_n, \tau)\},$$

where

$$\mathrm{EL}(z, w; \mathbf{y}_n, \tau) = \frac{1}{n\lambda(T)} \sum_{i=1}^{n} \lambda\{t \in T : z(t) \leq y_i(t) \leq z(t) + \tau(t)\} \mathbf{1}\{\forall_{t \in T}\ w(t) - \tau(t) \leq y_i(t) \leq z(t) + \tau(t)\},$$

$$\mathrm{HL}(z, w; \mathbf{y}_n, \tau) = \frac{1}{n\lambda(T)} \sum_{i=1}^{n} \lambda\{t \in T : w(t) - \tau(t) \leq y_i(t) \leq w(t)\} \mathbf{1}\{\forall_{t \in T}\ w(t) - \tau(t) \leq y_i(t) \leq z(t) + \tau(t)\}.$$

and the population version is given by $ls_{MHR}(x, y; Y, \tau) = \min\{\mathrm{EL}(z, w; Y, \tau), \mathrm{HL}(z, w; Y, \tau)\}$, where

$$\mathrm{EL}(z, w; Y, \tau) = \frac{1}{\lambda(T)} \mathrm{E}[\lambda\{t \in T : z(t) \leq Y(t) \leq z(t) + \tau(t)\} \mathbf{1}\{\forall_{t \in T}\ w(t) - \tau(t) \leq Y(t) \leq z(t) + \tau(t)\}],$$

$$\mathrm{HL}(z, w; Y, \tau) = \frac{1}{\lambda(T)} \mathrm{E}[\lambda\{t \in T : w(t) - \tau(t) \leq Y(t) \leq w(t)\} \mathbf{1}\{\forall_{t \in T}\ w(t) - \tau(t) \leq Y(t) \leq z(t) + \tau(t)\}].$$

Again, we have $ls_{MHR}(y, y; \mathbf{y}_n, \tau) = ld_{MHR}(y; \mathbf{y}_n, \tau)$ and $ls_{MHR}(x, y; \mathbf{y}_n, \tau) \leq \min\{ld_{MHR}(x; \mathbf{y}_n, \tau), ld_{MHR}(y; \mathbf{y}_n, \tau)\}$ for all $\tau > 0$, and the same for the population version.

There are several ways to construct dissimilarity/distance matrix $D = [d_{ij}]$ from similarity matrix $S = [s_{ij}]$. In our analysis we use the following transformation proposed by Gower [17]:

$$d_{ij} = (s_{ii} + s_{jj} - 2s_{ij})^{1/2}. \tag{1}$$

Once a dissimilarity matrix is available, hierarchical cluster analysis can be easily performed on the set of observed curves. In the next section, the local versions of these new similarity measures are illustrated.

5. Examples

In this section we illustrate the behavior of the local half-region depth and the local modified half-region depth and some comparison with their global version using the Berkeley Growth Study. A data set on wind speed is also analyzed using only the modified versions since half-region depth, and hence the local version, is constant due to the fact that all the curves have at least one intersection with the others. Further examples are available in the Online Supplement, where three other data sets are analyzed: handwritten vowels, individual household electric power consumption, and data from remote sensing devices.

5.1. Berkeley Growth Study

The Berkeley Growth Study [42] recorded the heights of 54 girls and 39 boys between the ages of 1 and 18 years. Heights were measured at 31 ages for each child, and the standard error of these measurements was about 3 mm, tending to be larger in early childhood and lower in later years. This functional data set is also studied, among others, in [36] and available in the R package `fda` under the name `growth`. This data set is considered a benckmark in classification problems and up-to-date reviews of the performance of depth based methods are available in [6, 30].

Figure 1 shows in light blue the height growth curve of boys and in light red that of the girls. We consider the clustering problem by ignoring the gender of the children and record how the local half-region similarity depth and the local modified half-region similarity depth are able to reconstruct the two groups. For this aim, after transforming the similarity matrix to a dissimilarity matrix using formula (1), a standard hierarchical clustering procedure is performed with the Ward1 agglomerative method [33]. The performance is evaluated by the mis-classification rate (MCR).

We first run a Monte Carlo experiment with 1000 replications by resampling without replacement 70 growth curves (approximately $3/4$ of the sample size) and evaluating the MCR for each value of τ corresponding to the percentile in the interval $[0.15, 1.00]$ of the empirical distribution of the sup distance between all pairs of curves. Figure 2 reports the MCR as a function of the percentile; all the reported curves were smoothed using LOWESS with smoother span equals to $1/5$. For the local half-region depth (left panel), the median (solid black) shows a minimum around 0.6 while for the local modified half-region depth (right panel), there is a plateau in the interval 0.2–0.4; 50% bands (dashed black) are large due to the small sample size of the data set. MCR for the whole data set is reported in red and for both methods the minimum is at a quantile order of 0.39.

We perform the analysis using this "optimal" quantile order; the confusion matrices are reported in Table 1 together with the MCR. In the same table, for comparison reason, we report the performance of a DD-classifier based on half-region depth and modified half-region depth. The clustering procedures are performing reasonably well with respect to the classification procedures. The local modified version is performing better with respect to the local version in the clustering procedures, while it is the opposite for the classification procedures; this is apparent in the dendrograms 1 and in the silhouette plots 2 reported in the Online Supplement.

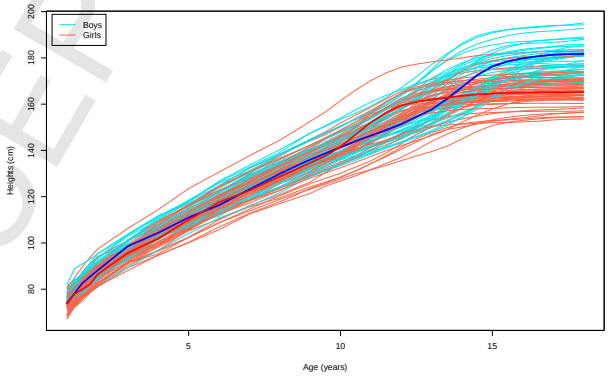

Figure 1: Berkeley Growth Study. In blue (boys) and red (girls) the deepest curves according to the modified half-region depth based on the two groups separately.

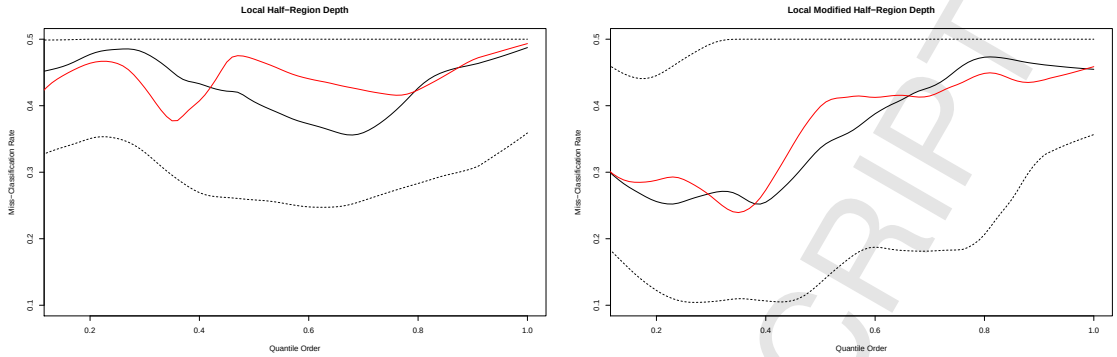

Figure 2: Berkeley Growth Study. Mis-classification rate versus quantile order of the empirical distribution of the sup distance between all pairs of curves. Black median, black dashed envelope of the 50%, red observed behavior for the complete data set. All the curves are smoothed.

Furthermore, Spearman's correlation coefficient between the local depth based on the whole data set and the depth based on one gender at time show high values: half-region 0.488 (boys), 0.757 (girls); modified half-region 0.584 (boys), 0.869 (girls) indicating that the local depth based on the whole data set is able to rank the units approximately in the same order as the depth based on one gender at time. The behavior of Spearman's correlation coefficient as a function of the percentile in the interval $[0.15, 1.00]$ of the empirical distribution of the sup distance between all pairs of curves is reported in Figure 3. Since, when considered separately, the two groups are unimodal, Spearman's correlation coefficient is always high especially for the local modified half-region depth.

	Clustering				Classification			
	LHR		LMHR		HR		MHR	
	Boy	Girl	Boy	Girl	Boy	Girl	Boy	Girl
Boy	39	0	39	0	32	7	23	16
Girl	32	22	17	41	6	48	21	33
MCR	0.34		0.18		0.14		0.40	

Table 1: Berkeley Growth Study data. Confusion matrix and mis-classification rate (MCD) by clustering and classification methods.

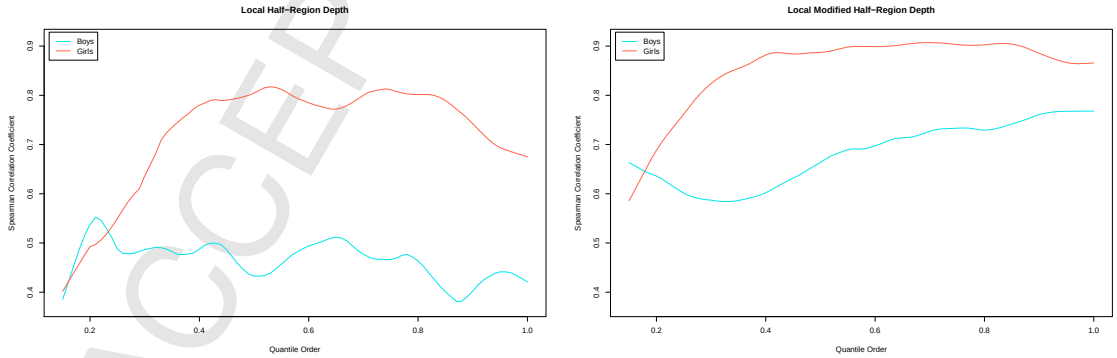

Figure 3: Berkeley Growth Study. Spearman's correlation coefficient between local depth based on the whole data set and depth based on one gender versus quantile order of the empirical distribution of the sup distance between all pairs of curves. All the curves are smoothed.

5.2. Wind speed

For this example we use a data set obtained from the Col De la Roa (Italy) meteorological station available at `intra.tesaf.unipd.it/Sanvito`. We concentrate on the wind speed recorded regularly every 15 minutes in the years 2001–2004. After removing 42 incomplete days, we obtain 1420 time series of length $4 \times 24 = 96$. We apply the modified half-region depth and the local modified half-region depth with $\tau = 2.619$, which corresponds to the 20% quantile order of the empirical distribution of the sup norm between two time series.

The two depths provide very different rankings for most series, while some ranks are preserved; see Figure 3 in the Online Supplement where we report the DD plot. This is confirmed by the plots in Figure 4, where each series is plotted with a different color and width according to its rank; darker and wider means higher rank. We also investigate the possible presence of subgroups or structures. To this end, we evaluate the dissimilarity measures based on the local modified half-region depth and modified half-region depth and run a hierarchical cluster analysis based on the Ward distance.

In Figure 5, the dendrograms and the silhouette graphs are reported. Inspection of the dendrograms and several silhouette plots (not reported) shows that two groups are suggested by the modified half-region depth while three groups are supported by the local version. The classification is very different. Groups based on depth dissimilarity are very similar and they do not identify different patterns (see Figure 6 in the Online Supplement), while the first group provided by local depth dissimilarity is characterized by days of low wind with modest changes during the day; the second group shows an interesting pattern during the day: low wind during the night and in the early morning with almost absence of wind between 8 am–10 am, moderate wind speed for the rest of the day until 9 pm. The third group is formed by days with higher wind speed, with variability during the day and peaks.

The first two groups correspond to the cold season (circular mean around December and mean resultant length of 0.52) and the hot season (circular mean around June and mean resultant length of 0.44) respectively, while the third group, given its nature, is spread over the whole year, with some predominance in the hot season (however, Watson's test for circular uniformity is strongly rejected with a p-value smaller than 0.01; the circular mean is around June and the mean resultant length is 0.18) as illustrated in Figure 6. Further comments and analysis are performed on this data set and reported in the Online Supplement.

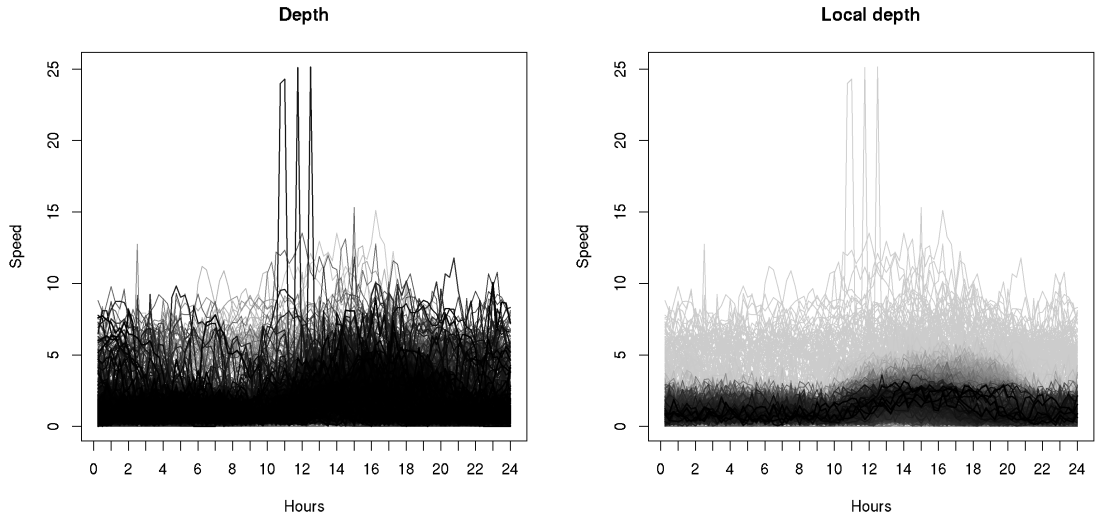

Figure 4: Wind Speed. Ranks provided by modified half-region depth (left panel) and local modified half-region depth (right panel); darker and wider means higher rank.

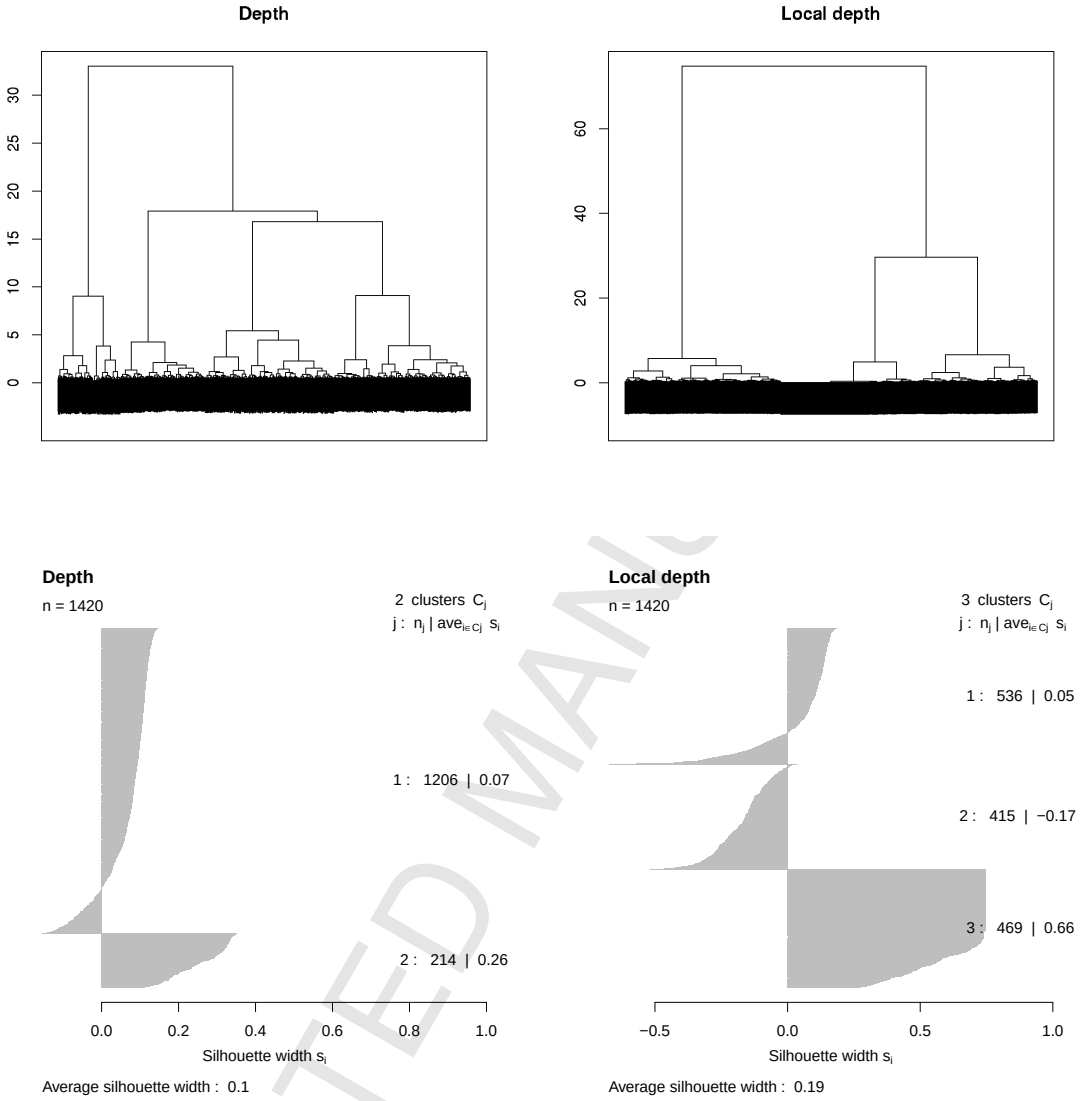

Figure 5: Wind Speed. First row, dendograms, second row silhouette plot, first column modified half-region depth, second column local modified half-region depth.

6. Monte Carlo experiments

To study the behavior of the newly introduced procedures, we ran a Monte Carlo experiment. Two groups of equal size $n \in \{50, 100\}$ were considered with trajectories generated according to the following model

$$Y_{i,t} = \mu_{i,t} + Z_t \quad i \in \{1, 2\},$$

where Z_t is a multivariate normal distribution, with the same distribution for both groups, with zero mean vector and variance and covariance matrix with elements $\gamma(s,t) = k \exp\{-c|t-s|^m\}$, k, c, m strictly greater than zero. Here, c and m are fixed to 1 and 1.75, respectively, while k changes in the range $[0.1, 1]$ with steps of 0.1 size, $t \in [0, 1]$ in a grid of 30 equally spaced points, $\mu_{1,t} = 4t$ and $\mu_{2,t} = 4t + d$, where $d \in \{0, \ldots, 20\}$. Figure 26 in the Online Supplement reports a sample for the cases $k \in \{0.1, 0.5, 1\}$ and $d \in \{0, 1, 2\}$ and sample size $n = 50$. For each combination

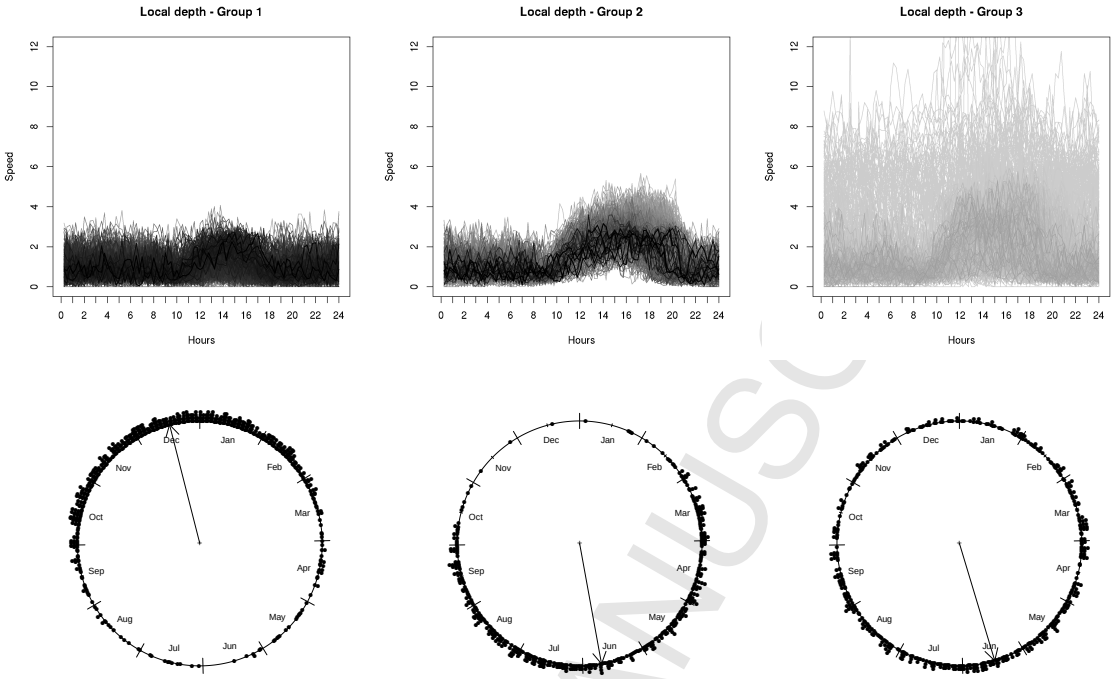

Figure 6: Wind Speed. Groups provided by a cluster analysis based local modified half-region depth similarity when three groups are formed. Top: Curves are plotted with color and thickness according to their depth/local depth. Bottom: Classification of the days according to the three groups.

of the experiment settings, we ran $R = 500$ Monte Carlo replications. This setting is similar to the one proposed, e.g., in [28]; see also [44]. We computed the half-region depth, the modified half-region depth, and their local versions. For the local versions we considered a grid of values for τ that corresponds to the following quantile orders $0.10, 0.15, \ldots, 0.40, 0.45, 0.46, 0.47, \ldots, 0.54, 0.55, 0.60, \ldots, 0.95, 1.00$. Let $\mathbf{y}_{i,j}$ be the jth Monte Carlo replication from the ith group of size n; we let $(\mathbf{y}_{1,j}, \mathbf{y}_{2,j})$ denote the joined dataset, r stand for the function that provides the ranks of a given vector argument, and ρ be Pearson's correlation coefficient. We evaluated the performance of our procedures by two indices based on Spearman's correlation, viz.

$$\mathrm{corr}_i(\tau) = \frac{1}{R}\sum_{j=1}^{R} \rho[r\{ld(\mathbf{y}_{i,j}; (\mathbf{y}_{1,j}, \mathbf{y}_{2,j}), \tau)\}, r\{d(\mathbf{y}_{i,j}, \mathbf{y}_{i,j})\}], \quad i \in \{1, 2\},$$

which is a measure of how good is the local depth based on the whole sample at reproducing the ranks of the depth based on one of the group data sets and

$$\mathrm{corr}(\tau) = \frac{1}{R}\sum_{j=1}^{R} \rho[r[ld\{(\mathbf{y}_{1,j}, \mathbf{y}_{2,j}); (\mathbf{y}_{1,j}, \mathbf{y}_{2,j}), \tau\}] - r[d\{(\mathbf{y}_{1,j}, \mathbf{y}_{2,j}), (\mathbf{y}_{1,j}, \mathbf{y}_{2,j})\}]],$$

which is the agreement between the ranks of the whole data set given by the local depth and the depth.

Figures 27 and 28 in the Online Supplement report $\mathrm{corr}_1(\tau)$ for local half-region depth and local modified half-region depth, respectively; $\mathrm{corr}_2(\tau)$ has very similar behavior and is not reported, Figures 29 and 30 in the Online Supplement report $\mathrm{corr}(\tau)$ for local half-region depth and local modified half-region depth, respectively. There are no significant differences in the behavior of the local half-region depth and the local modified half-region depth in this Monte Carlo experiment. As is expected for this configuration $\mathrm{corr}_1(\tau)$ is large for a quantile order smaller than 0.5,

with a maximum around 0.2–0.3 for small d and a value of almost 1 for quantile order in the interval 0.45–0.50 for large value of d. As k increases, a large value of d is needed in order to get a large value of Spearman's correlation coefficient $\mathrm{corr}_1(\tau)$. For quantile order greater than 0.5, Spearman's correlation coefficient $\mathrm{corr}_1(\tau)$ is around 0 as soon as d is large. Furthermore, considering the behavior of $\mathrm{corr}(\tau)$ shows that the correlation is very high (almost always close to 1) for quantile order of τ greater than 0.5 while the index approaches 0 as d is large and the quantile order is small.

7. Concluding remarks

Statistical data depth provides useful tools for the analysis of complex datasets such as functional data. Given the nature of depth functions, however, an analysis based on this concept may not be well suited for non homogeneous data, where subgroups or substructures are present. In contrast, local depth is able to capture local features of the data and to provide a better analysis of the observations. Local modified half-region depth proves to be a suitable tool for the analysis of functional data as well as high-dimensional data. Hierarchical clustering can be easily performed using similarity measures based on local modified half-region depth; classification is also straightforward but not investigated in this work.

Acknowledgments. The author is grateful to the editor, associate editor, and two anonymous referees for their comments that have largely contributed to improve the original manuscript. The author is grateful to Prof. Mario Romanazzi for his comments on an earlier version of the manuscript and to Domenico Toniolo, Orazio Agostinelli and Sante Bertin for their comments on a very earlier version of the manuscript. All statistical analyses were performed on SCSCF (`www.dais.unive.it/scscf`), a multiprocessor cluster system owned by Ca' Foscari University of Venice running under GNU/Linux and on HPC (`icts.unitn.it/risorse-di-calcolo`), a multiprocessor cluster system owned by the University of Trento running under GNU/Linux.

References

[1] C. Agostinelli, M. Romanazzi, Local depth of multidimensional data, Working Paper 2008.3, Dipartimento di Statistica, Università Ca' Foscari, Venezia, Italy, 2008.
[2] C. Agostinelli, M. Romanazzi, Local depth, J. Statist. Plann. Inf. 141 (2011) 817–830.
[3] C. Agostinelli and R. Rotondi. Analysis of macroseismic fields using statistical data depth functions: Considerations leading to attenuation probabilistic modelling, *Bull. Earthquake Eng.*, pp. 1–16, 2015.
[4] G. Aneiros, E.G. Bongiorno, R. Cao, P. Vieu, Eds., Functional Statistics and Related Fields, Springer, New York, 2017.
[5] A. Chakraborty, P. Chaudhuri, On data depth in infinite dimensional spaces, Ann. Inst. Statist. Math. 66 (2014) 303–324.
[6] A. Cuesta, M. Febrero-Bande, M. Oviedo de la Fuente, The dd^g-classifier in the functional setting, TEST 26 (2017) 119–142.
[7] J.A. Cuesta-Albertos, A. Nieto-Reyes, The random Tukey depth, Comput. Statist. Data Anal. 52 (2008) 4979–4988.
[8] A. Cuevas, A partial overview of the theory of statistics with functional data, J. Statist. Plann. Inf. 147 (2014) 1–23.
[9] A. Cuevas, M. Febrero-Bande, R. Fraiman, Robust estimation and classification for functional data via projection-based depth notions, Comput. Statist. 22 (2007) 481–496.
[10] J. Demongeot, A. Hamie, A. Laksaci, M. Rachdi, Relative-error prediction in nonparametric functional statistics: Theory and practice, J. Multivariate Anal. 146 (2016) 261–268.
[11] Y. Ding, X. Dang, H. Peng, D. Wilkins, Robust clustering in high dimensional data using statistical depths, BMC Bioinformatics 8 (2007).
[12] B. Everitt, Cluster Analysis, Heinemann Educational Books, London, 1974.
[13] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer, New York, 2006.
[14] R. Fraiman, G. Muniz, Trimmed means for functional data, TEST 10 (2001) 419–440.
[15] A. Goia, P. Vieu, An introduction to recent advances in high/infinite dimensional statistics, J. Multivariate Anal. 146 (2016) 1–6.
[16] W. González-Manteiga, P. Vieu, Statistics for functional data, Comput. Statist. Data Anal. 51 (2007) 4788–4792.
[17] J.C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (1966) 325–338.
[18] J.A. Hartigan, Clustering Algorithms, Wiley, New York, 1975.
[19] L. Horváth, P. Kokoszka, Inference for Functional Data With Applications, Springer, New York, 2012.
[20] A. Inselberg, The plane with parallel coordinates, The Visual Computer 1 (1985) 69–91.
[21] R. Jörnsten, Clustering and classification based on the L_1 data depth, J. Multivariate Anal. 90 (2004) 67–89.
[22] L.-Z. Kara, A. Laksaci, M. Rachdi, P. Vieu, Data-driven kNN estimation in nonparametric functional data analysis, J. Multivariate Anal. 153 (2017) 176–188.
[23] P. Kokoszka, M. Reimherr, Introduction to Functional Data Analysis, Chapman and Hall/CRC, London, 2017.
[24] M.R. Kosorok, Introduction to Empirical Processes and Semiparametric Inference, Springer, Berlin, 2008.
[25] S. Kuhnt, A. Rehage, An angle-based multivariate functional pseudo-depth for shape outlier detection, J. Multivariate Anal. 146 (2016) 325–340.

[26] R.Y. Liu, On a notion of data depth based on random simplices, Ann. Statist. 18 (1990) 405–414.
[27] R.Y. Liu, R.J. Serfling, D.L. Souvaine, Data Depth: Robust Multivariate Analysis, Computational Geometry, and Applications, AMS Bookstore, 2006.
[28] S. López-Pintado, J. Romo, On the concept of depth for functional data, J. Amer. Statist. Assoc. 104 (2009) 718–734.
[29] S. López-Pintado, J. Romo, A half-region depth for functional data, Comput. Statist. Data Anal. 55 (2011) 1679–1695.
[30] K. Mosler, P. Mozharovsky, Fast DD-classification of functional data, Statist. Papers, in press.
[31] K. Mosler, Y. Polyakova, General notions of depth for functional data, arXiv:1208.1981, 2012.
[32] R. Moustafa, E. Wegman, Multivariate continuous data – parallel coordinates, In: Graphics of Large Datasets, Statistics and Computing, pp. 143–155, Springer New York, 2006.
[33] F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? J. Classif. 31 (2014) 274–295.
[34] S. Nagy, An overview of consistency results for depth functionals, In G. Aneiros, E.G. Bongiorno, R. Cao, P. Vieu, Eds., Functional Statistics and Related Fields, Chap. 25, Springer, New York, 2017.
[35] A. Nieto-Reyes, H. Battery Statistical functional depth, In: G. Aneiros, E.G. Bongiorno, R. Cao, P. Vieu, Eds., Functional Statistics and Related Fields, Chap. 26, Springer, New York, 2017.
[36] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer, New York, 2002.
[37] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, 2006.
[38] R.J. Serfling, Generalized quantile processes based on multivariate depth functions, with applications in nonparametric multivariate analysis, J. Multivariate Anal. 83 (2002) 232–247.
[39] C. Sguera, P. Galeano, R. Lillo, Spatial depth-based classification for functional data, TEST 23 (2014) 725–750.
[40] H. Tailen, R. Eubank, Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators, Wiley, Chichester, West Sussex, 2015.
[41] Y. Tian, Y.R. Gel, Fast community detection in complex networks with a *k*-depths classifier, In: S. Ejaz Ahmed, Ed., *Big and Complex Data Analysis. Methodologies and Applications*, pp. 139–157, Springer, 2017.
[42] R.D. Tuddenham, M.M. Snyder, Physical growth of California boys and girls from birth to eighteen years, U. of California Publications in Child Development 1 (1954) 183–364.
[43] J.W. Tukey, Mathematics and picturing of data, In: Proceedings of International Congress of Mathematics 2 (1975) 523–531.
[44] A.T.A. Wood, G. Chan, Simulation of stationary Gaussian processes in $C[0, 1]$, J. Comput. Graph. Statist. 3 (1994) 409–432.
[45] Y. Zuo, R.J. Serfling, General notions of statistical depth function, Ann. Statist. 28 (2000) 461–482.