A half-region depth for functional data

Sara López-Pintado [a,b,*], Juan Romo [c]

[a] Department of Biostatistics, Columbia University, NY, USA
[b] Departamento de Economía, Métodos Cuantitativos e Historia Económica, Universidad Pablo de Olavide, Sevilla, Spain
[c] Departamento de Estadística, Universidad Carlos III de Madrid, Madrid, Spain

ARTICLE INFO

Article history:
Received 7 December 2009
Received in revised form 27 October 2010
Accepted 27 October 2010
Available online 31 October 2010

Keywords:
Functional data
Data depth
Order statistics
High-dimensional data

ABSTRACT

A new definition of depth for functional observations is introduced based on the notion of "half-region" determined by a curve. The half-region depth provides a simple and natural criterion to measure the centrality of a function within a sample of curves. It has computational advantages relative to other concepts of depth previously proposed in the literature which makes it applicable to the analysis of high-dimensional data. Based on this depth a sample of curves can be ordered from the center-outward and order statistics can be defined. The properties of the half-region depth, such as consistency and uniform convergence, are established. A simulation study shows the robustness of this new definition of depth when the curves are contaminated. Finally, real data examples are analyzed.

1. Introduction

Steadily increasing attention is being paid to the analysis of functional data in recent years (see Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; González-Manteiga and Vieu, 2007). A fundamental task in functional data analysis is to provide a natural ordering within a sample of curves, which makes it possible to define ranks and L-statistics. In this paper we introduce a new definition of depth for functional observations based on the concepts of "hypograph" and "epigraph" of a curve. This functional depth provides a criterion for ordering the sample of curves from center-outward. The notion of statistical depth was first analyzed for multivariate observations, and many different definitions of depth have been studied in the literature, for example Mahalanobis (1936), Tukey (1975), Oja (1983), Liu (1990), Donoho and Gasko (1992), Liu et al. (1999), Zuo and Serfling (2000) and Zuo (2003). Most of these multivariate depths are not adequate for high-dimensional data, therefore their applicability is restricted to low-dimensional vector observations. Recently, alternative notions of depth for functional data have been introduced which can be adapted to high-dimensional data without a large computational burden (see Fraiman and Muniz, 2001; Cuevas et al., 2006, 2007; Cuesta-Albertos and Nieto-Reyes, 2008; López-Pintado and Jörnsten, 2007 and López-Pintado and Romo, 2009). In this paper we propose an alternative graph-based notion of depth which is simple, computationally fast, and can be easily adapted to high-dimensional data.

This paper is organized as follows. The new half-region depth S_H is defined in Section 2. In Section 3 we analyze the finite-dimensional version of S_H and prove some properties such as consistency and uniform convergence. We extend these results to the infinite-dimensional case in Section 4. Section 5 deals with a modified version of S_H which is more convenient for irregular functional data. In Section 6 the half-region depths are compared to other proposed depths using simulated curves from different contaminated models. Real data examples are analyzed in Section 7.

* Corresponding address: Department of Biostatistics, Columbia University, 722W, 168th Street, NY, USA. Tel.: +1 212 305 2271.
E-mail addresses: sl2929@columbia.edu, sloppin@upo.es (S. López-Pintado).

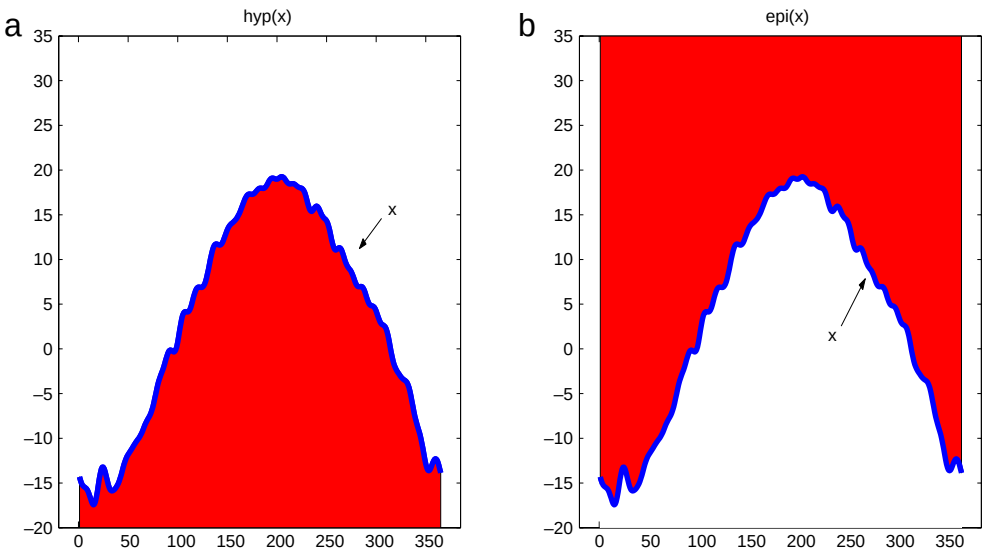

Fig. 1. (a) Hypograph and (b) epigraph of function x.

2. Half-region depth

Let $C(I)$ be the space of continuous functions defined on a compact interval I. This is a Banach space with the supremum norm. Consider a stochastic process X with sample paths in $C(I)$ with distribution P. Let $x_1(t), x_2(t), \ldots, x_n(t)$ be a sample of curves from P. The graph of a function x in $C(I)$ will be denoted as $G(x)$, thus

$$G(x) = \{(t, x(t)),\ t \in I\}.$$

Define the hypograph *(hyp)* and the epigraph *(epi)* of a function x in $C(I)$ as

$$hyp(x) = \{(t, y) \in I \times \mathbb{R} : y \leq x(t)\},$$
$$epi(x) = \{(t, y) \in I \times \mathbb{R} : y \geq x(t)\}.$$

In Fig. 1 the hypograph and epigraph of a function x are represented. The half-region depth is defined as follows.

Definition 1. The half-region depth at x with respect to a set of functions $x_1(t), \ldots, x_n(t)$ is

$$S_{n,H}(x) = \min\{G_{1n}(x), G_{2n}(x)\},$$

where

$$\begin{aligned} G_{1n}(x) &= \frac{\sum_{i=1}^{n} I(G(x_i) \subset hyp(x))}{n} \\ &= \frac{\sum_{i=1}^{n} I(x_i(t) \leq x(t),\ t \in I)}{n}, \\ G_{2n}(x) &= \frac{\sum_{i=1}^{n} I(G(x_i) \subset epi(x))}{n} \\ &= \frac{\sum_{i=1}^{n} I(x_i(t) \geq x(t),\ t \in I)}{n}, \end{aligned}$$

and $I(A)$ is the indicator function of the set A.

Hence, the half-region sample depth at x is the minimum between the proportion of functions of the sample whose graph is in the hypograph of x and the corresponding proportion for the epigraph of x.

The population version of $S_{n,H}(x)$ is

$$S_H(x) = \min\{G_1(x), G_2(x)\},$$

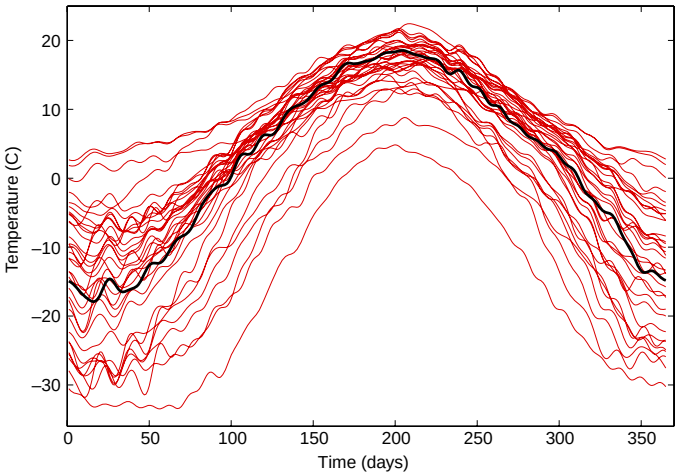

Fig. 2. Daily temperatures during one year in 35 weather stations in Canada; the curve which maximizes the depth $S_{n,H}$ is represented in dark black.

where

$$G_1(x) = P(G(X) \subset hyp(x)) = P(X(t) \leq x(t),\ t \in I)$$

and

$$G_2(x) = P(G(X) \subset epi(x)) = P(X(t) \geq x(t),\ t \in I).$$

The symmetry of these expressions provides an alternative way of defining the half-region depth at a point x with respect to P,

$$S_H(x) = \min\{DG_1(x), DG_2(x)\},$$

where

$$DG_1(x) = P(G(x) \subset hyp(X))$$

and

$$DG_2(x) = P(G(x) \subset epi(X)).$$

The sample version of this second way of defining the half-region depth is obtained by substituting P by the empirical distribution P_n and coincides with the one proposed in Definition 1.

A deepest curve, or S_H-sample median $\widehat{\tau}_n$, is a curve from the sample which maximizes the half-region depth,

$$\widehat{\tau}_n = \underset{x \in \{x_1, \ldots, x_n\}}{\arg\max}\ S_{n,H}(x),$$

and the S_H-population median is defined as a curve in $C(I)$ which maximizes S_H. Moreover, if the sample of curves $x_1, x_2, \ldots, x_n$ are ordered according to decreasing values of $S_{n,H}(x_i)$, we obtain order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(1)}$ denotes the deepest observation and $x_{(n)}$ the shallowest one. Based on this ordering we can also define the α-trimmed mean as the average of the proportion $1 - \alpha$ deepest curves. Fig. 2 shows a real data example that consists of the daily temperatures during one year in 35 different weather stations in Canada. The curve represented in dark black color is the deepest one. This curve can be used to illustrate the representative pattern within the sample of curves.

In López-Pintado and Romo (2007) different inference tools (such as the central region, the scale curve, and a rank test) were introduced based on the ordering provided by the band depth (see López-Pintado and Romo, 2009 for more details on this depth). All these methods can also be applied using the half-region depth presented here. In particular, the p-central region is the band delimited by the proportion p of deepest curves. To illustrate this idea, in Fig. 3(a) we represent the central region defined by the 15% deepest curves (based on $S_{n,H}$) for the daily temperatures example described in the previous paragraph. The scale curve, $A(p)$, is defined as

$$A(p) = \text{area}\{(t, y) \in I \times \mathbf{R},\ \min_{i=1,\ldots,\lfloor np \rfloor} x_{(i)}(t) \leq y \leq \max_{i=1,\ldots,\lfloor np \rfloor} x_{(i)}(t)\},$$

where $\lfloor np \rfloor$ is the nearest integer greater than or equal to np, and it measures how the area of the band determined by the fraction p of deepest curves increases with p. The scale curve is a simple tool that can be used for visualizing the center-outward dispersion of a sample of curves. It can also be used as an exploratory plot for comparing dispersion in different groups of curves and for detecting outliers. For example, an abrupt increase in the scale curve could indicate the presence of an outlier. The scale curve for the temperature function sample is plotted in Fig. 3(b).

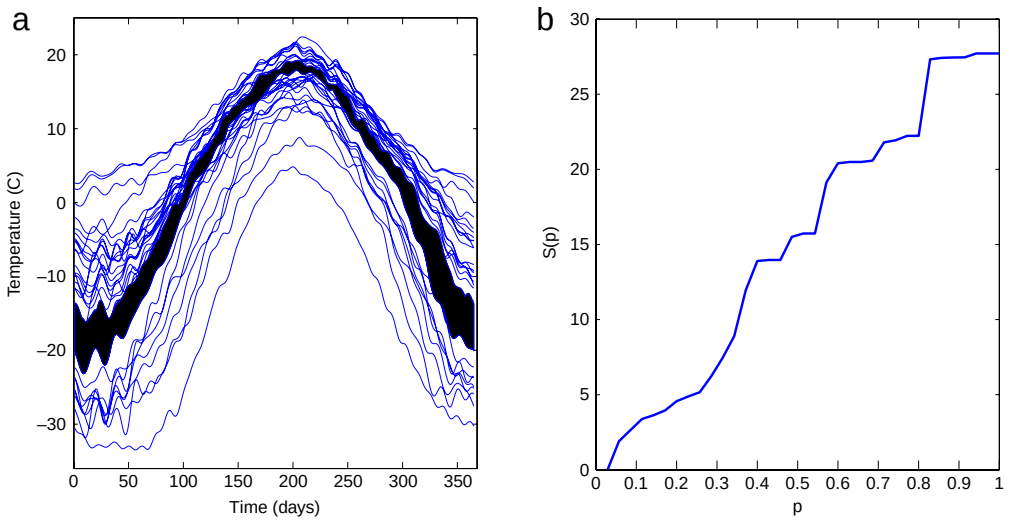

Fig. 3. (a) Trimmed region determined by the five deepest curves from the sample of functions (based on $S_{n,H}$) representing the daily temperature during one year in 35 weather stations in Canada and (b) the scale curve for this sample of curves.

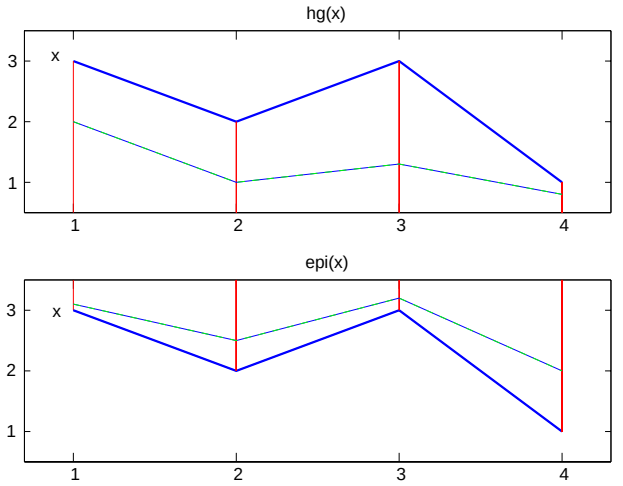

Fig. 4. Hypograph and epigraph of a point in $\mathbb{R}^4$ using parallel coordinates. The points represented with dashed lines in the top and low panels of the figure belong to the hypograph and the epigraph, respectively.

3. Finite-dimensional version

The concepts of hypograph and epigraph described in the previous section can be easily adapted to finite-dimensional data. Denote by $x(k)$ the kth component of the d-dimensional vector x. Consider each point in $\mathbb{R}^d$ as a function defined in the set of indices $\{1, \ldots, d\}$, the hypograph and epigraph of a point $x = (x(1), x(2), \ldots, x(d))$ can be expressed respectively as

$$hyp(x) = \{(k, y) \in \{1, 2, \ldots, d\} \times \mathbb{R} : \ y \leq x(k)\}$$

and

$$epi(x) = \{(k, y) \in \{1, 2, \ldots, d\} \times \mathbb{R} : \ y \geq x(k)\}.$$

A convenient way of representing a d-dimensional vector is using parallel coordinates (see Inselberg, 1985 and Wegman, 1990), where the d axes are now parallel and equidistant, and the coordinates of the vector are represented as points on these axes connected by straightlines. Fig. 4 gives the hypograph and the epigraph of a point $x = (3, 2, 3, 1) \in \mathbb{R}^4$ using parallel coordinates. In addition, two points belonging to $hyp(x)$ and $epi(x)$ are represented with dashed lines. An alternative interpretation can be obtained using the cartesian representation of the points in $\mathbb{R}^d$(with $d \leq 3$). Fig. 5 shows the hypograph (lower quadrant) and the epigraph (upper quadrant) of a point x in $\mathbb{R}^2$ using its representation in cartesian coordinates.

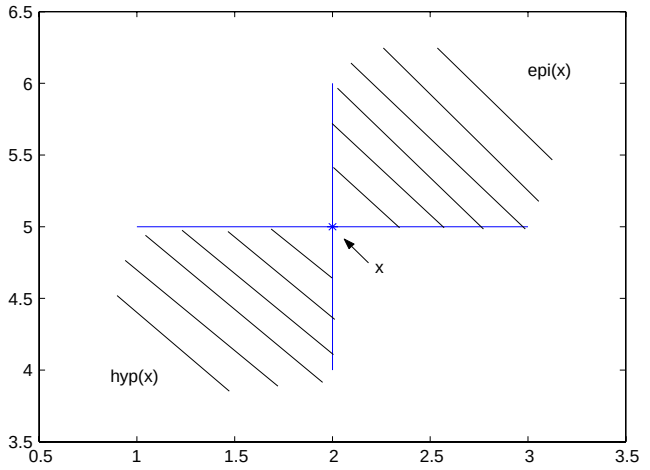

Fig. 5. Representation in cartesian coordinates of the hypograph and the epigraph of a point $x \in \mathbb{R}^2$.

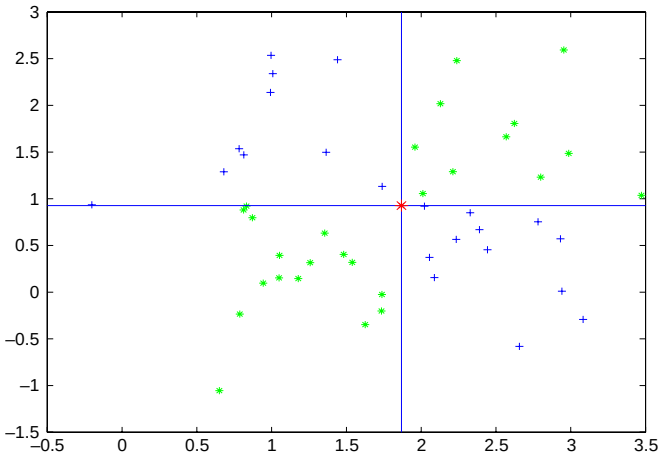

Fig. 6. Deepest point from a sample of 50 normal observations.

Let X be a d-dimensional random variable with distribution function F. $X \leq x$ and $X \geq x$ are the notation for $\{X(k) \leq x(k), k = 1, \ldots, d\}$ and $\{X(k) \geq x(k),\ k = 1, \ldots, d\}$, respectively. If we particularize the half-region depth to the finite-dimensional case, we obtain

$$\begin{aligned} S_H(x, F) &= S_H(x) = \min\{P(X \leq x), P(X \geq x)\} \\ &= \min\{F_X(x), F_{-X}(-x)\} = \min\{F_X(x),\ F_Y(y)\}, \end{aligned}$$

where $Y = -X$ and $y = -x$.

Let $x_1, \ldots, x_n$ be a random sample from the variable X, the sample version of the half-region depth is

$$\begin{aligned} S_{n,H}(x) &= \min \left\{ \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}, \frac{\sum_{i=1}^{n} I(x_i \geq x)}{n} \right\} \\ &= \min\{F_{X_n}(x), F_{Y_n}(y)\}. \end{aligned}$$

Fig. 6 shows 50 points simulated from a normal bivariate distribution and it illustrates the way of computing the half-region depth $S_{n,H}$ of the deepest point from the sample represented with an asterisk. The proportion of data in the upper right quadrangle is 12/50 and in the lower left quadrangle is 17/50; therefore, the half-region depth of the red point is 12/50.

The main advantage of the half-region depth relative to other multivariate depths is that it is fast to compute and applicable to high-dimensional data (where $n \ll d$). It can be easily shown that the computational cost of the half-region depth of a point in $\mathbb{R}^d$ with respect to a sample of n d-dimensional points is $O(n \cdot d)$. In what follows we describe the theoretical properties satisfied by S_H.

The half-region depth is invariant with respect to translations and some types of dilations. Let A be a positive (or negative) definite diagonal matrix and $b \in \mathbb{R}^d$, then

$$S_H(Ax + b, F_{Ax+b}) = S_H(x, F).$$

In the following propositions we establish some other properties of this notion of depth.

Proposition 1. *For $d = 1$ the half-region depth $S_H(x)$ can be expressed as*

$$\begin{aligned} S_H(x) &= \min\{P(X \le x), 1 - P(X < x)\} \\ &= \min\{F(x), 1 - F(x^-)\}, \end{aligned}$$

and is equivalent to Tukey's half-space depth. Moreover, the value that maximizes S_H is the usual median in $\mathbb{R}$.

The half-region depth decreases to zero when the point tends to infinity.

Proposition 2 (*Vanishing at Infinity*)**.** *Let $x \in \mathbb{R}^d$,*

$$\sup_{\|x\| \ge M} S_H(x) \longrightarrow 0, \quad \text{when } M \to \infty,$$

and

$$\sup_{\|x\| \ge M} S_{n,H}(x) \xrightarrow{a.s.} 0, \quad \text{when } M \to \infty.$$

Note that the previous proposition implies that

$$S_H(x) \longrightarrow 0, \quad \text{when } \|x\|_\infty \to \infty,$$
$$S_{n,H}(x) \xrightarrow{a.s.} 0, \quad \text{when } \|x\|_\infty \to \infty.$$

Proposition 3. *$S_H(\cdot)$ is an upper semicontinuous function. Moreover, if F is absolutely continuous then $S_H(\cdot)$ is continuous.*

The proofs of Propositions 2 and 3 are postponed to next section since they are particular cases of the same properties in the functional case. In the next proposition we establish the uniform convergence of $S_{n,H}$ to its population version.

Proposition 4. *$S_{n,H}$ is uniformly consistent:*

$$\sup_{x \in \mathbb{R}^d} |S_{n,H}(x) - S_H(x)| \xrightarrow{a.s.} 0, \quad \text{when } n \to \infty.$$

Moreover, if $S_H(x)$ is uniquely maximized at τ and τ_n is a sequence of random variables with $S_{n,H}(\tau_n) = \sup_{x \in \mathbb{R}^d} S_{n,H}(x)$, then

$$\tau_n \xrightarrow{a.s.} \tau, \quad \text{when } n \to \infty.$$

Proof. Applying Glivenko–Cantelli's theorem in $\mathbb{R}^d$, we have that

$$\sup_{x \in \mathbb{R}^d} |F_{Xn}(x) - F_X(x)| \xrightarrow{a.s.} 0, \quad \text{when } n \to \infty,$$

and

$$\sup_{y \in \mathbb{R}^d} |F_{Yn}(y) - F_Y(y)| \xrightarrow{a.s.} 0, \quad \text{when } n \to \infty.$$

Therefore,

$$\sup_{x \in \mathbb{R}^d} |\min\{F_{X_n}(x), F_{Y_n}(y)\} - \min\{F_X(x), F_Y(y)\}| \le \sup_{x \in \mathbb{R}^d} |F_{X_n}(x) - F_X(x)| + \sup_{y \in \mathbb{R}^d} |F_{Y_n}(y) - F_Y(y)| \xrightarrow{a.s.} 0.$$

The second part of the theorem is proven using arguments similar to the ones proposed by Arcones et al. (1994) to show the consistency of the simplicial median. By Proposition 3, S_H is an upper semicontinuous function, then, $\limsup_{n \to \infty} S_H(y_n) \le S_H(y)$, if $y_n \underset{n \to \infty}{\to} y$. Also, using that $\lim_{\|x\| \to \infty} S_H(x) = 0$, and that $S_H(x)$ is uniquely maximized at τ, we have that for every $\varepsilon > 0$, $S_H(\tau) - \sup_{|x - \tau| \ge \varepsilon} S_H(x) > 0$. Hence, for the following argument consider $\delta = S_H(\tau) - \sup_{|x-\tau| \ge \varepsilon} S_H(x) > 0$.

To prove that $\tau_n \xrightarrow{a.s.} \tau$, it is sufficient to establish that

$$P\{\sup_{n \ge l} |\tau_n - \tau| > \varepsilon\} \longrightarrow 0, \quad \text{when } l \to \infty.$$

Recall that

$$\begin{aligned}
P\{\sup_{n\geq l}|\tau_n-\tau|>\varepsilon\} &\leq P\{\sup_{n\geq l}(S_H(\tau)-S_H(\tau_n))\geq\delta\}\\
&\leq P\{(\sup_{n\geq l}(S_H(\tau)-S_{n,H}(\tau))+\sup_{n\geq l}(S_{n,H}(\tau_n)-S_H(\tau_n)))\geq\delta\}\\
&\leq P\{\sup_{n\geq l}(S_H(\tau)-S_{n,H}(\tau))\geq\delta/2\}+P\{\sup_{n\geq l}(S_{n,H}(\tau_n)-S_H(\tau_n))\geq\delta/2\}\\
&\leq P\{\sup_{n\geq l}\sup_x|S_H(x)-S_{n,H}(x)|\geq\delta/2\}+P\{\sup_{n\geq l}\sup_x|S_{n,H}(x)-S_H(x)|\geq\delta/2\}\\
&\leq 2P\{\sup_{n\geq l}\sup_x|S_{n,H}(x)-S_H(x)|\geq\delta/2\}\underset{l\to\infty}{\longrightarrow}0.
\end{aligned}$$

Therefore, $P\{\sup_{n\geq l}|\tau_n-\tau|>\varepsilon\}\longrightarrow 0$ when $l\to\infty$. □

4. Properties of the functional half-region depth

Here, we extend some of the properties established in the previous section to the functional version of the half-region depth S_H. Let $x_1,\ldots,x_n$ be independent copies of a stochastic process X in $C(I)$ with distribution function P. Assume that the stochastic process X is tight, i.e.,

$$P(\|X\|_\infty\geq M)\longrightarrow 0,\quad \text{when } M\to\infty. \tag{1}$$

The depth satisfies a linear invariance property. Consider a and b functions in $C(I)$, where $a(t)>0$ or $a(t)<0$ for every $t\in I$. Then

$$S_H(x,P_X)=S_H(ax+b,P_{aX+b}).$$

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

Proposition 5. *The depths S_H and $S_{n,H}$ satisfy that*

$$\sup_{\|x\|_\infty\geq M}S_H(x)\longrightarrow 0,\quad \textit{when } M\to\infty, \tag{2}$$

and

$$\sup_{\|x\|_\infty\geq M}S_{n,H}(x)\xrightarrow{a.s.}0,\quad \textit{when } M\to\infty. \tag{3}$$

Proof. The quantity $\sup_{\|x\|_\infty\geq M}S_H(x)$ can be decomposed depending on where the supremum is achieved in the following way:

$$\sup_{\|x\|_\infty\geq M}S_H(x)\leq\sup_{\|x\|_\infty\geq M\cap\|x\|_\infty=\sup x(t)}S_H(x)+\sup_{\|x\|_\infty\geq M\cap\|x\|_\infty=\sup(-x(t))}S_H(x).$$

Now,

$$\begin{aligned}
\sup_{\|x\|_\infty\geq M\cap\|x\|_\infty=\sup x(t)}S_H(x) &\leq \sup_{\|x\|_\infty\geq M\cap\|x\|_\infty=\sup x(t)}P(X(t)\geq x(t),\ t\in I)\\
&\leq \sup_{\|x\|_\infty\geq M\cap\|x\|_\infty=\sup x(t)}P(\|X\|_\infty\geq\|x\|_\infty)\\
&\leq P(\|X\|_\infty\geq M)\to 0,\quad \text{when } M\to\infty.
\end{aligned}$$

And also,

$$\begin{aligned}
\sup_{\|x\|_\infty\geq M\cap\|x\|_\infty=\sup(-x(t))}S_H(x) &\leq \sup_{\|x\|_\infty\geq M\cap\|x\|_\infty=\sup(-x(t))}P(X(t)\leq x(t),\ t\in I)\\
&\leq \sup_{\|x\|_\infty\geq M\cap\|x\|_\infty=\sup(-x(t))}P(-X(t)\geq -x(t),\ t\in I)\\
&\leq \sup_{\|x\|_\infty\geq M\cap\|x\|_\infty=\sup(-x(t))}P(\|-X\|_\infty\geq\|x\|_\infty)\\
&\leq P(\|-X\|_\infty\geq M)\to 0,\quad \text{when } M\to\infty.
\end{aligned}$$

To prove that $S_{n,H}$ converges almost surely to zero we use the same decomposition as before. Hence, here we just present a sketch of the proof. If $\|x\|_\infty = \sup x(t)$,

$$\begin{aligned}\sup_{\|x\|_\infty \geq M \cap \|x\|_\infty = \sup(x(t))} S_{n,H}(x) &\leq \sup_{\|x\|_\infty \geq M \cap \|x\|_\infty = \sup x(t)} \frac{1}{n}\sum_{i=1}^{n} I\{X_i(t) \geq x(t),\ t \in I\} \\ &\leq \sup_{\|x\|_\infty \geq M \cap \|x\|_\infty = \sup x(t)} \frac{1}{n}\sum_{i=1}^{n} I\{\|X_i\|_\infty \geq \|x\|_\infty\} \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \sup_{\|x\|_\infty \geq M} I\{\|X_i\|_\infty \geq \|x\|_\infty\}.\end{aligned}$$

In what follows we show that $X_M = \sup_{\|x\|_\infty \geq M} I\{\|X_i\|_\infty \geq \|x\|_\infty\}$ converges almost surely to 0 when M tends to infinity. Define $Y_M = I\{\|X_i\|_\infty \geq M\}$, since

$$0 \leq X_M \leq Y_M,$$

it is sufficient to prove that $Y_M \xrightarrow{a.s.} 0$, or equivalently that

$$P(\sup_{M \geq l} I\{\|X_i\|_\infty \geq M\} > \varepsilon) \longrightarrow 0, \quad \text{when } l \to \infty.$$

It is easy to see that the following inequality holds,

$$\sup_{M \geq l} I\{\|X_i\|_\infty \geq M\} \leq I\{\|X_i\|_\infty \geq l\},$$

and it implies that

$$\begin{aligned}P(\sup_{M \geq l} I\{\|X_i\|_\infty \geq M\} > \varepsilon) &\leq P(I\{\|X_i\|_\infty \geq l\} > \varepsilon) \\ &= P(\|X_i\|_\infty \geq l) \longrightarrow 0, \quad \text{when } l \to \infty.\end{aligned}$$

Thus, we have proven that $X_M \xrightarrow{a.s.} 0$, when $M \to \infty$. In the case that $\|x\|_\infty = \sup(-x(t))$ the proof is analogous. □

Proposition 6. *$S_H(\cdot)$ is an upper semicontinuous functional. Moreover, if P has absolutely continuous marginals, then $S_H(\cdot)$ is continuous.*

Proof. To prove that $S_H(\cdot)$ is upper semicontinuous we show that $\limsup_{n\to\infty} S_H(y_n) \leq S_H(y)$, when $y_n \xrightarrow{\|\cdot\|_\infty} y$.

$$\begin{aligned}\limsup_{n\to\infty} S_H(y_n) &= \limsup_{n\to\infty} \min\{G_1(y_n), G_2(y_n)\} \\ &= \limsup_{n\to\infty} \min\{P(G(X) \subset hyp(y_n)), P(G(X) \subset epi(y_n))\} \\ &\leq \min\{\limsup_{n\to\infty} P(G(X) \subset hyp(y_n)), \limsup_{n\to\infty} P(G(X) \subset epi(y_n))\} \\ &\leq \min\{P(G(X) \subset hyp(y)), P(G(X) \subset epi(y))\} = S_H(y).\end{aligned}$$

To establish the continuity of the functional $S_H(\cdot)$ in $C(I)$ with respect to the supremum norm it is sufficient to prove that both $G_1(\cdot)$ and $G_2(\cdot)$ are continuous. In what follows we prove that $G_1(\cdot)$ is continuous; the case $G_2(\cdot)$ is analogous. We have to see that if $x_n \xrightarrow{\|\cdot\|_\infty} x$ then $|G_1(x_n) - G_1(x)| \underset{n\to\infty}{\longrightarrow} 0$. Recall that

$$\begin{aligned}|G_1(x_n) - G_1(x)| &= |P(G(X) \subset hyp(x_n)) - P(G(X) \subset hyp(x))| \\ &\leq P(G(X) \subset hyp(x_n) \cap G(X) \subsetneq hyp(x)) + P(G(X) \subsetneq hyp(x_n) \cap G(X) \subset hyp(x)).\end{aligned}$$

Considering that the marginals of the distribution P are continuous it is easy to prove that

$$\begin{aligned}&P(G(X) \subset hyp(x_n) \cap G(X) \subsetneq hyp(x)) \underset{n\to\infty}{\longrightarrow} 0 \quad \text{and} \\ &P(G(X) \subsetneq hyp(x_n) \cap G(X) \subset hyp(x)) \underset{n\to\infty}{\longrightarrow} 0.\end{aligned} \tag{4}$$

Hence, G_1 is a continuous function. □

In the next theorem we establish the strong consistency of the sample half-region depth. To facilitate reading, we use the notation $X_i \leq x$ and $X_i \geq x$ to denote the events $\{X_i(t) \leq x(t),\ t \in I\}$ and $\{X_i(t) \geq x(t),\ t \in I\}$, respectively.

Theorem 1. *$S_{n,H}$ is strongly consistent,*

$$S_{n,H}(x) \xrightarrow{a.s.} S_H(x).$$

Proof. The sample half-region depth $S_{n,H}(x)$ can be expressed as

$$S_{n,H}(x) = \min\left(\frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\}, \frac{1}{n}\sum_{i=1}^{n} I\{X_i \geq x\}\right).$$

By the law of large numbers and the continuity of the minimum,

$$\min\left(\frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\}, \frac{1}{n}\sum_{i=1}^{n} I\{X_i \geq x\}\right) \xrightarrow{a.s} \min(P(X \leq x), P(X \geq x)),$$

and then $S_{n,H}(x) \xrightarrow{a.s.} S_H(x)$. □

Finally, we establish the uniform consistency of $S_{n,H}$ and the strong consistency of the argument that maximizes $S_{n,H}$. The half-region depth can be expressed as a transformation of two empirical processes. We first present some notation; see, e.g., Pollard (1984). Let f be a measurable functional from $C(I)$ to $\mathbb{R}$. The value $P_n f$ is the expectation of f under the empirical distribution and Pf is the expectation of f based on P:

$$P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

and

$$Pf = \int f(X)\mathrm{d}P.$$

For a subset E of $C(I)$, consider the family of functions $\mathcal{F}_1$

$$\mathcal{F}_1 = \{f_x^{(1)} : x \in E\}, \tag{5}$$

where $f_x^{(1)} : E \subset C(I) \longrightarrow \{0, 1\}$ is defined as

$$f_x^{(1)}(y) = I\{x(t) \leq y(t),\ t \in I\}.$$

Therefore,

$$f_x^{(1)}(X_i) = \begin{cases} 1, & \text{if } x(t) \leq X_i(t), \text{ for every } t \in I, \\ 0, & \text{in any other case.} \end{cases}$$

Analogously, define

$$\mathcal{F}_2 = \{f_x^{(2)} : x \in E\}, \tag{6}$$

where $f_x^{(2)} : E \subset C(I) \longrightarrow \{0, 1\}$,

$$f_x^{(2)}(y) = I\{x(t) \geq y(t),\ t \in I\}.$$

Hence,

$$f_x^{(2)}(X_i) = \begin{cases} 1, & \text{if } x(t) \geq X_i(t), \text{ for every } t \in I, \\ 0, & \text{in any other case.} \end{cases}$$

The following theorem provides the strong uniform consistency of the half-region depth for classes of functions $\mathcal{F}_1$ and $\mathcal{F}_2$ with finite bracketing number.

Theorem 2. *If the classes of functions $\mathcal{F}_1$ and $\mathcal{F}_2$ defined in (5) and (6) have finite bracketing number ($\mathcal{N}_{[\,]}(\varepsilon, \mathcal{F}_1, L_1(P)) < \infty$, $\mathcal{N}_{[\,]}(\varepsilon, \mathcal{F}_2, L_1(P)) < \infty$) for every $\varepsilon > 0$, then*

$$\sup_{x \in E} |S_{n,H}(x) - S_H(x)| \xrightarrow{a.s.} 0.$$

Proof. The result is a consequence of the Glivenko–Cantelli Theorem (for example in Van Der Vaart, 1998) and the following equation,

$$\sup_{x \in E} |S_{n,H}(x) - S_H(x)| \leq \sup_{f \in \mathcal{F}_1} |P_n f - Pf| + \sup_{f \in \mathcal{F}_2} |P_n f - Pf|.$$

Since $\mathcal{F}_1$ and $\mathcal{F}_2$ have finite bracketing number, $\sup_{x \in E} |S_{n,H}(x) - S_H(x)| \xrightarrow{a.s.} 0$. □

We now provide some examples of families of functions $\mathcal{F}$ that satisfy the condition of finite bracketing number. In addition, a continuity condition for the probability distribution is needed.

C1 Given $\varepsilon > 0$, there exists $\gamma > 0$, such that for every pair of functions $z_i, z_j \in C(I)$ if $\|z_i - z_j\|_\infty \leq \gamma$ then $P(z_j \leq X \leq z_i) \leq \varepsilon$.

Definition 2. A subset E of $C(I)$ is equicontinuous if for each $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that for every $x \in E$ and for every $t, s \in I$,

$$\text{if } |t - s| < \delta \quad \text{then } |x(t) - x(s)| < \varepsilon.$$

In the next theorem we establish the uniform convergence of $S_{n,H}(x)$ to its population version over the set of functions E.

Theorem 3. *If $E \subset C(I)$ is equicontinuous and P is a probability distribution in $C(I)$ satisfying condition* C1*, then $S_{n,H}(x)$ is uniformly consistent at E:*

$$\sup_{x \in E} |S_{n,H}(x) - S_H(x)| \xrightarrow{a.s} 0, \quad \textit{when } n \to \infty. \tag{7}$$

Proof. Without loss of generality, assume that $I = [0, 1]$. The following decomposition holds

$$\sup_{x \in E} |S_{n,H}(x) - S_H(x)| \leq \sup_{x \in E,\ \|x\|_\infty \leq M} |S_{n,H}(x) - S_H(x)| + \sup_{x \in E,\ \|x\|_\infty \geq M} |S_{n,H}(x) - S_H(x)|.$$

Since the second term converges almost surely to zero when M tends to infinity by Proposition 5, given M sufficiently large, we just need to prove that

$$\sup_{x \in E,\ \|x\|_\infty \leq M} |S_{n,H}(x) - S_H(x)| \xrightarrow{a.s.} 0, \quad \text{when } n \to \infty. \tag{8}$$

We have that

$$\sup_{x \in E_M} |S_{n,H}(x) - S_H(x)| \leq \sup_{x \in E_M} \left| \frac{1}{n} \sum_{i=1}^{n} I(x \leq X_i) - P(x \leq X) \right| + \sup_{x \in E_M} \left| \frac{1}{n} \sum_{i=1}^{n} I(x \geq X_i) - P(x \geq X) \right|,$$

where $E_M = \{x \in E : \|x\|_\infty \leq M\}$. If we consider the family of functions $\mathcal{F}_1^M = \{f_x^{(1)} : x \in E_M\}$ and $\mathcal{F}_2^M = \{f_x^{(2)} : x \in E_M\}$, then

$$\sup_{x \in E_M} |S_{n,H}(x) - S_H(x)| \leq \sup_{f \in \mathcal{F}_1^M} |P_n f - Pf| + \sup_{f \in \mathcal{F}_2^M} |P_n f - Pf|.$$

Therefore, it is sufficient to establish that the families $\mathcal{F}_1^M$ and $\mathcal{F}_2^M$ have finite bracketing number ($\mathcal{N}_{[]}(\varepsilon, \mathcal{F}_1^M, L_1(P)) < \infty$ and $\mathcal{N}_{[]}(\varepsilon, \mathcal{F}_2^M, L_1(P)) < \infty$). We will only prove it for $\mathcal{F}_1^M$, because the case $\mathcal{F}_2^M$ is analogous. Given $\varepsilon > 0$, we need to construct a finite number of functions $z_1, \ldots, z_p$ that determine brackets, $[f_{z_j}, f_{z_i}]$, covering the family $\mathcal{F}_1^M$ and satisfying

$$P(f_{z_j} - f_{z_i}) = P(z_j \leq X \leq z_i) < \varepsilon. \tag{9}$$

Since the condition C1 is satisfied, there exists $\nu > 0$ such that if $\|z_i - z_j\|_\infty < \nu$ then (9) holds. By the equicontinuity of E_M, given $\nu > 0$, there exists $\delta > 0$, such that if $|s - t| < \delta$ then $|x(s) - x(t)| < \nu$ for every $x \in E_M$. Consider the set of functions defined as constants in the intervals $[0, \delta)$, $[\delta, 2\delta)$, $[2\delta, 3\delta)$, $\ldots$, $[([1/\delta] - 1)\delta, 1)$, and taking values in the sequence: $-[M/\nu]\nu, \ldots -\nu, 0, \nu, 2\nu, \ldots, [M/\nu]\nu$. The total number p of possible functions that can be constructed like this is finite and we denote them as $z_1, \ldots, z_p$. Define the following set of indicator functions $f_z : C \to \{0, 1\}$,

$$\{f_{z_k},\ k \in \{1, \ldots, p\}\} = \{f_{z_k}(X) = I\{z_k(t) \leq X(t),\ t \in [0, 1]\} :\ k \in \{1, \ldots, p\}\}.$$

This set of functions allows us to construct ε-brackets that cover the family $\mathcal{F}_1^M$: for every $f_x \in \mathcal{F}_1^M$, two functions z_i and z_j can be chosen, such that

$$P(f_{z_j} - f_{z_i}) < \varepsilon,$$

and $f_{z_i} \leq f_x \leq f_{z_j}$. The functions z_i and z_j are chosen to satisfy $z_j(t) \leq x(t) \leq z_i(t),\ t \in [0, 1]$, and $\sup_{t \in [0,1]} |z_i(t) - z_j(t)| \leq \nu$. This implies that

$$I_{\{z_i \leq X\}} \leq I_{\{x \leq X\}} \leq I_{\{z_j \leq X\}},$$
$$f_{z_i} \leq f_x \leq f_{z_j},$$

and

$$\begin{aligned} P(f_{z_j} - f_{z_i}) &= P(z_j \leq X) - P(z_i \leq X) \\ &= P(z_j \leq X < z_i) < \varepsilon. \end{aligned}$$

Hence, the family of functions $\mathcal{F}_1^M$ and $\mathcal{F}_2^M$ have finite bracketing number and the result in (8) holds. □

The following theorem proves the uniform convergence of the value that maximizes $S_{n,H}$.

Theorem 4. *Let P be a distribution satisfying condition* C1 *in the equicontinuous set E. If $S_H(\cdot)$ is uniquely maximized at $\tau \in E$ and τ_n is a sequence of functions in E with $S_{n,H}(\tau_n) = \sup_{x\in E} S_{n,H}(x)$ then*

$$\tau_n \xrightarrow{a.s.} \tau, \quad \text{when } n \to \infty. \tag{10}$$

Proof. We have to show that

$$P(\sup_{n\geq l} \|\tau_n - \tau\|_\infty \geq \varepsilon) \xrightarrow[l\to\infty]{} 0.$$

$(E, \|\cdot\|_\infty)$ is a metric space and $S_H(\cdot)$ is upper semicontinuous in E and satisfies

$$\sup_{\|x\|_\infty \geq M,\ x\in E} S_H(x) \xrightarrow[M\to\infty]{} 0.$$

Then the proof is analogous to the one in Proposition 4. □

The set of functions $Lip_{\alpha,A}(I)$ given by

$$Lip_{\alpha,A}(I) = \{x : I \to \mathbb{R},\ \text{such that } |x(t_1) - x(t_2)| \leq A|t_1 - t_2|^\alpha,\ \text{for every } t_1,\ t_2 \in I\},$$

is equicontinuous and, therefore, it satisfies Theorem 3. Hence, the sample half-region depth $S_{n,H}$ converges uniformly to S_H over the set $Lip_{\alpha,A}(I)$.

5. A modified half-region depth

Here we introduce a modified version of the half-region depth, less restrictive than the definition described before, that can be used for the analysis of irregular (non-smooth) curves which could crossover a lot. This new depth is based on what we denote as the superior (SL) and the inferior (IL) lengths, which are defined by:

$$SL(x) = \frac{1}{\lambda(I)} E[\lambda\{t \in I : x(t) \leq X(t)\}],$$
$$IL(x) = \frac{1}{\lambda(I)} E[\lambda\{t \in I : x(t) \geq X(t)\}],$$

where λ stands for the Lebesgue measure on $\mathbb{R}$. $SL(x)$ can be interpreted as the "proportion of time" that the stochastic process X is greater than x. Similarly, $IL(x)$ is the "proportion of time" that the process X is smaller than x. The modified half-region depth at x is:

$$MS_H(x) = \min\{SL(x), IL(x)\}.$$

Let $x_1, \ldots, x_n$ be a set of curves with distribution P. The sample version of the notion of depth is obtained by substituting P by the empirical distribution P_n,

$$MS_{n,H}(x) = \min\{SL_n(x), IL_n(x)\},$$

where

$$SL_n(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \leq x_i(t)\}$$

and

$$IL_n(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \geq x_i(t)\}.$$

Some properties of the modified half-region depth can be derived following similar arguments to the ones used for the half-region depth. Proposition 1 holds trivially, whereas, the vanishing at infinity property in Proposition 2 is not satisfied for the modified half-region depth. The consistency results in Theorems 1 and 2 are satisfied and the proofs are analogous to the ones for S_H. For Theorem 2, the functions in $\mathcal{F}_1$ and $\mathcal{F}_2$ are now defined as $f_{1,x}(X) = \lambda\{t \in I, x(t) \leq X(t)\}$ and $f_{2,x}(X) = \lambda\{t \in I, x(t) \geq X(t)\}$, respectively.

6. Simulation results

In this section we report the results of a simulation study where the robustness of the half-region and modified half-region depths are analyzed. We simulated curves from different contaminated models and compared the performance of several estimators of the population mean of the generated curves. In particular, we compare trimmed mean estimates based

on S_H and MS_H with those obtained using the band depth (BD) and the modified band depth (MBD) introduced in López-Pintado and Romo (2009). In addition, we include in the comparison the trimmed mean based on the L1-depth introduced originally for multivariate data in Zuo and Serfling (2000). The ordering provided by this depth is based on the L1-distances to the sample points and can be easily extended to functional data (see Ferraty and Vieu, 2006). We have also included in the comparison the coordinate-wise median function, defined as the univariate median at each point t, and sample mean. We first consider four models (M1–M4) with what we call "magnitude" contamination. These models were analyzed by Fraiman and Muniz (2001) and López-Pintado and Romo (2009). They all consist in adding some outliers to an elementary model M0 defined as

$$X_i(t) = f(t) + e_i(t), \quad 1 \le i \le n,$$

where $t \in [0, 1]$, n is the number of curves generated, $e_i(t)$ is a Gaussian stochastic process with zero mean and covariance function

$$E(e_i(t)e_i(s)) = \exp\{-|t - s|\},$$

and the function $f(t) = 4t$.

The asymmetric total contamination model (M1) is defined by

$$Y_i(t) = X_i(t) + \epsilon_i M, \quad 1 \le i \le n,$$

where ϵ_i takes values 1 with probability q and 0 with probability $1 - q$. The constant M is the contamination size.

A model of symmetric contamination (M2) can be obtained in the following way:

$$Y_i(t) = X_i(t) + \epsilon_i \sigma_i M, \quad 1 \le i \le n,$$

where ϵ_i and M are defined as in the previous model and σ_i is a sequence of random variables independent of ϵ_i taking values 1 and -1 with probability 1/2.

A partially contaminated model (M3) can be expressed as follows:

$$Y_i(t) = X_i(t) + \epsilon_i \sigma_i M, \quad \text{for } t \ge T_i \; 1 \le i \le n, \quad \text{and}$$
$$Y_i(t) = X_i(t), \quad \text{for } t < T_i,$$

where ϵ_i, M and σ_i are defined as in model 2 and T_i is a random number generated from a uniform distribution on (0, 1).

The fourth model considered here is a peak contamination model (M4) expressed as:

$$Y_i(t) = X_i(t) + \epsilon_i \sigma_i M, \quad \text{for } T_i \le t \le T_i + l, \; 1 \le i \le n, \quad \text{and}$$
$$Y_i(t) = X_i(t), \quad \text{for } t \notin [T_i, T_i + l],$$

where $l = 2/30$ and T_i is a random number from a uniform distribution on $[0, 1 - l]$. The idea behind this model is to contaminate the curves only in a short interval.

In addition to these four models we have considered a family of models (models 5–9 in Table 2) where the type of contamination is very different to previous ones. More specifically, the contamination is in "shape" instead of "magnitude". Thus, we have included outliers which are not far away from the rest of the curves in terms of any distance but differ in shape. We use a family of models to generate shape outliers which was introduced in López-Pintado and Romo (2009) based on the covariance kernels presented in Wood and Chan (1994). More concretely, the curves are generated from a Gaussian process with covariance kernel $\gamma(s, t) = k \exp\{-c|t - s|^{\mu}\}$, with $s, t \in [0, 1]$, and $k, c, \mu > 0$. Parameters k, c and μ control shape. For example, when c is increased, the generated functions are more irregular. However, when μ and k are increased the generated curves become smoother. The shape contaminated models in Table 2 are defined as a mixture of a basic model $X_i(t) = f(t) + e_{1i}(t)$, and another model with the same mean, $Y_i(t) = f(t) + e_{2i}(t)$, where $1 \le i \le n$. In particular, $f(t) = 4t$ for models 5 and 6, $f(t) = 4t^2$ for models 7 and 8 and $f(t) = 4t^3$ for model 9. The error term $e_{i1}(t)$ is a Gaussian stochastic process with zero mean and covariance function $\gamma_1(s, t) = \exp\{-|t - s|^2\}$ and $e_{i2}(t)$ is a Gaussian process with zero mean and covariance function $\gamma_2(s, t) = k_2 \exp\{-c_2|t - s|^{\mu_2}\}$, with values k_2, c_2 and μ_2 chosen to generate more irregular curves. For simplicity, in models 5–9 we have fixed $c_2 = 1$ and $k_2 = 1$ and changed the value of μ (e.g., in model 5, $\mu_2 = 0.2$, $c_2 = 1$ and $k_2 = 1$). The contaminated models are given by $Z_i(t) = (1 - \varepsilon)X_i(t) + \varepsilon Y_i(t)$, $1 \le i \le n$, where ε is a Bernoulli variable $Be(q)$ and q is a small contamination probability. Essentially, we contaminate a sample of smooth curves from $X_i(t)$ with curves from $Y_i(t)$ having different covariance functions and providing more irregular curves.

For each model we have considered $N = 500$ replications for $n = 50$ curves and different possible estimators of f: half-region and modified half-region trimmed mean, band depth and modified band depth trimmed mean, coordinate-wise median, L1-depth trimmed mean (introduced in Ferraty and Vieu, 2006), and sample mean. The trimming level is always equal to $\alpha = 0.2$.

For each of the N replications the integrated error is calculated and evaluated at $L = 30$ equally spaced points in [0, 1],

$$EI(j) = \frac{1}{L} \sum_{k=1}^{L} [\widehat{g}_n(k/L) - f(k/L)]^2,$$

Table 1
$N = 500$, $n = 50$, $q = 0.1$, $\alpha = 0.2$, $M = 5$ and $M = 25$.

		M0	M1	M2	M3	M4
$M = 5$	S_H	**0.0243** (**0.0071**)	0.3415 (0.3431)	0.0411 (0.0664)	0.0551 (0.0434)	0.0312 (0.0305)
	MS_H	0.0248 (0.0073)	**0.0605** (**0.0896**)	**0.0263** (**0.0079**)	0.0281 (0.0098)	0.0328 (0.0305)
	BD	0.0244 (0.0071)	0. 3045 (0.2987)	0.0412 (0.0699)	0.0547 (0.0423)	**0.0280** (**0.0278**)
	MBD	0.0259 (0.0079)	0.0669 (0.0963)	**0.0266** (**0.0085**)	**0.0265** (**0.0078**)	0.0314 (0.0286)
	Cm	0.0304 (0.0083)	0.0613 (0.0606)	0.0357 (0.0303)	0.0335 (0.0099)	0.0291 (0.0232)
	L1D	0.0263 (0.0079)	**0.0339** (**0.0405**)	0.0267 (0.008)	**0.0261** (**0.0075**)	0.0303 (0.0305)
	Mean	**0.0198** (**0.0058**)	0.3313 (0.2785)	0.0706 (0.0795)	0.0437 (0.0320)	**0.0235** (**0.0209**)
$M = 25$	S_H	**0.0246** (**0.0075**)	4.6040 (5.5241)	0.3486 (1.4232)	0.8385 (1.1873)	0.1417 (0.0953)
	MS_H	0.0251 (0.0073)	0.3372 (1.0982)	**0.0303** (**0.0407**)	0.1014 (0.1298)	0.1951 (0.1107)
	BD	**0.0246** (**0.0075**)	3.8420 (4.9652)	0.3437 (1.4175)	0.7711 (1.1190)	0.1036 (0.0822)
	MBD	0.0259 (0.0074)	0.4160 (1.1904)	**0.0302** (**0.0406**)	0.0462 (0.0381)	0.1794 (0.1059)
	Cm	0.0304 (0.0088)	**0.0559** (**0.0561**)	0.0390 (0.0134)	**0.0339** (**0.0106**)	**0.0310** (**0.0263**)
	L1D	0.0261 (0.0076)	**0.0335** (**0.0420**)	0.0305 (0.0404)	**0.0266** (**0.0080**)	**0.0285** (**0.0302**)
	Mean	**0.0199** (**0.0060**)	7.3518 (5.5890)	1.2284 (1.8041)	0.6613 (0.8589)	0.1495 (0.0820)

where $\widehat{g}_n$ denotes any of the estimators we proposed earlier. To compare the performance of these estimators in the different models, we calculate the mean integrated error and its standard deviation defined respectively as:

$$E = \frac{1}{N}\sum_{j=1}^{N} EI(j)$$

and

$$S = \left(\frac{1}{N}\sum_{j=1}^{N}(EI(j) - E)^2\right)^{1/2}.$$

Table 1 provides the results of the simulation study based on the first four models. The mean integrated error and its standard deviation (in parenthesis) are calculated for each estimation method and model with parameters $N = 500$, $n = 50$, $q = 0.1$, $M = 5$ and $M = 25$. When $M = 5$ the mean integrated error in the model with no contamination (M0) is minimized by the sample mean followed by the trimmed mean based on the half-region depth. Also in Model 4, where the contamination is only on a short interval of time the best performance is given by the mean followed by the band depth trimmed mean. For models M1 and M3 the best methods are the trimmed means based on the modified half-region depth and the L1-depth, whereas in M2 the mean integrated error is minimized by the modified band depth trimmed mean followed by the modified half-region trimmed mean. When the contamination constant is $M = 25$, the minimum mean integrated errors for models M1, M2 and M4 are obtained with the coordinate-wise median followed by the L1-depth trimmed mean. For Model 2 the mean integrated error is minimized with the modified half-region depth and modified band depth. Table 1 shows that there is no single method outperforming all the others in every contamination scheme presented in this first set of models (M1–M4). The best location estimator is going to depend strongly on the type of curves and outliers presented in the particular data set.

Table 2 contains the simulation results for the family of shape contaminated models (models 5–9) with 500 replications, $n = 50$ curves, $\alpha = 0.2$ and contamination probability $q = 0.15$. We have considered different values of μ_2 to change the shape of the curves generated from the process $Y(t)$. For simplicity, in all these models $k_2 = c_2 = 1$. As with previous models, we compare the mean, the coordinate-wise median, and the depth based trimmed means in terms of robustness. The mean integrated squared error using the different location estimators are represented for each shape contamination

Table 2
$N = 500, n = 50, q = 0.15$ and $\alpha = 0.2$.

		M5 ($\mu_2 = 0.2$)	M6 ($\mu_2 = 0.1$)	M7 ($\mu_2 = 0.2$)	M8 ($\mu_2 = 0.1$)	M9 ($\mu_2 = 0.1$)
$q = 0.15$	S_H	**0.0435** (**0.0470**)	**0.0454** (**0.0496**)	**0.0435** (**0.0484**)	**0.0417** (**0.0396**)	**0.0448** (**0.0453**)
	MS_H	0.0512 (0.0544)	0.0640 (0.0614)	0.0592 (0.0627)	0.0599 (0.0524)	0.0610 (0.0543)
	BD	**0.04** (**0.0439**)	**0.041** (**0.0464**)	**0.0397** (**0.0433**)	**0.0388** (**0.0366**)	**0.0407** (**0.0408**)
	MBD	0.0543 (0.0520)	0.0564 (0.0544)	0.0524 (0.0550)	0.0543 (0.0489)	0.0553 (0.0496)
	Cm	0.0491 (0.0421)	0.0488 (0.0449)	0.0481 (0.0475)	0.0463 (0.0405)	0.0499 (0.0435)
	$L1D$	0.0518 (0.0521)	0.0512 (0.0513)	0.0474 (0.0473)	0.0489 (0.0476)	0.05 (0.0476)
	$Mean$	0.0450 (0.0446)	0.0462 (0.0442)	0.0437 (0.0468)	0.0450 (0.0407)	0.0456 (0.0397)

Table 3
Percentage of times that an outlier is detected with 100 replications, 50 curves, and $\alpha = 0.2$.

	$\mu_2 = 0.1$ $k_2 = 1$	$\mu_2 = 0.2$ $k_2 = 1$	$\mu_2 = 0.3$ $k_2 = 1$	$\mu_2 = 0.1$ $k_2 = 2$	$\mu_2 = 0.2$ $k_2 = 2$	$\mu_2 = 0.3$ $k_2 = 2$
S_H	100	100	97	100	100	100
MS_H	6	19	4	18	16	18
BD	100	100	100	100	100	100
MBD	15	23	23	52	52	48
$L1D$	35	40	37	89	87	86

model (models 5–9). The minimum mean integrated squared errors always correspond to the trimmed mean based on the band depth BD and on the half-region depth S_H. This is due to the contamination type: shape more than magnitude. Thus, the band depth and the half-region depth perform well in terms of robustness with respect to these kinds of outliers because they assign low depth to curves with a different shape. However, the trimmed means based on other depths and the coordinate-wise median are less robust against shape contamination, since most of the contaminated curves can still be very central in the sample (in terms of distance) although the curve behaviour is different from the remaining functions. Therefore, when there is shape contamination in the data, the most robust location estimators are based on the band depth and the half-region depth.

The idea of functional depth can also be used to detect outliers within a sample of curves. To further explore this application of depth we have generated curves from models similar to those in Table 2 and compared the performance of different depths. We simulated 49 curves from the basic model X and one curve from the contaminated model Y using different values of the parameter μ_2 with $k_2 = c_2 = 1$. We replicated the simulation 100 times and counted the number of times the outlying curve is within the 20% least deepest curves from the sample. The results depend strongly on the notion of depth used (see Table 3). For the band depth BD and half-region depth S_H, the outlier was detected 100% of the times in almost all the models, whereas for the other notions of depth this percentage varies but was always significantly lower. We conclude that detecting shape contaminated curves is not an easy task and the half-region depth and band depth performed significantly better compared to the other functional depths. In addition, the half-region depth is very easy to compute and does not depend on any exogenous parameter.

7. Real data examples

7.1. Functional data examples

In this section we apply the half-region and modified half-region depth to real data examples, most of them also analyzed in Ramsay and Silverman (2005). The first example, already introduced in Section 2, consists of the temperature recorded in a year from 35 different weather stations in Canada (Fig. 7). These curves have been smoothed using a Fourier basis. The five deepest curves based on the half-region depth and modified half-region depth are represented in dark black in Fig. 7(a) and (b), respectively. They illustrate the behaviour of the most central observations within the group of curves.

A second real data example consists of the growth curves of a sample of 54 girls (Fig. 8). In this case the curves were smoothed using a spline basis. The deepest curves obtained using the half-region depth and modified half-region depth are highlighted in thick black and show the representative growth pattern within the sample of curves. These depth based order

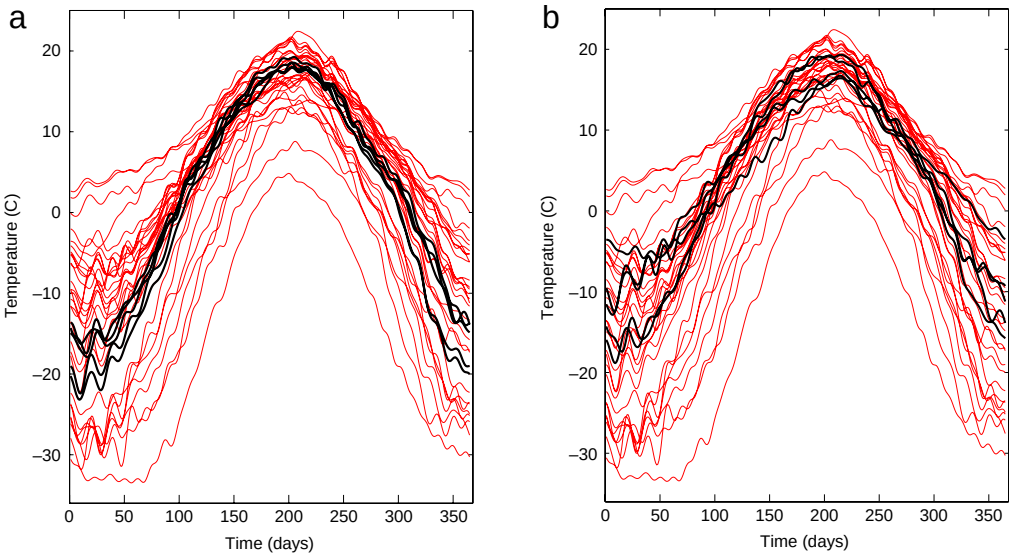

Fig. 7. Temperatures in different weather stations in Canada during one year. The curves were smoothed using a Fourier basis. The thick black curves represent the deepest curves using (a) the half-region depth and (b) the modified half-region depth.

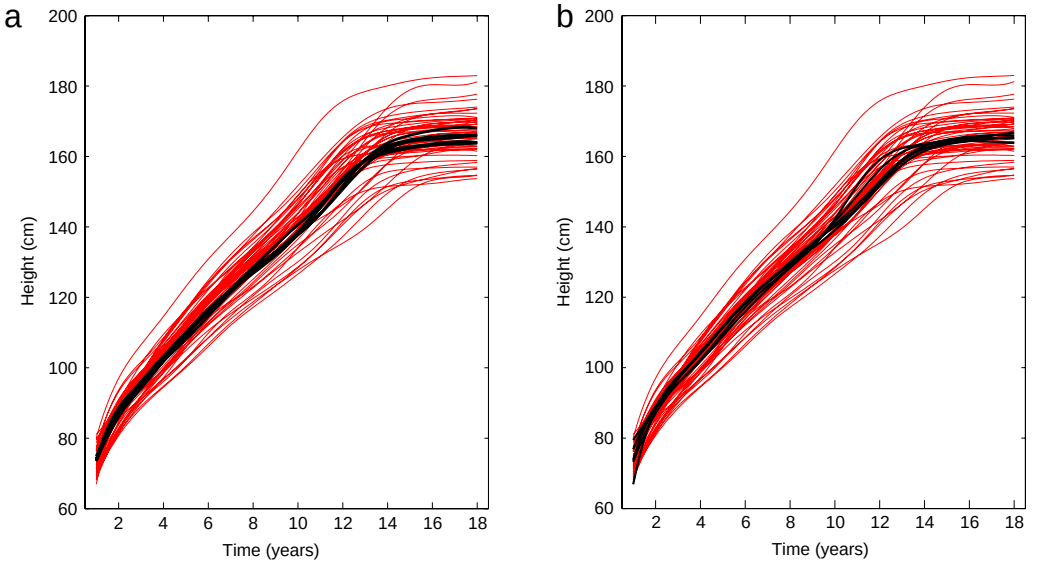

Fig. 8. Heights in cm of 54 girls during their first 18 years of life. The original curves were smoothed using a spline basis. In dark black we have represented the deepest curves (a) based on the half-region depth and (b) based on the modified half-region depth.

statistics can be used as a starting point for extending robust statistical methods to functional data. For instance, a functional version of k-medians could be introduced treating the deepest function of a group as the median curve.

In the third and last functional data example we have applied the half-region depth and the modified half-region depth to analyze how the relative diameter of a sample of Laricio trees changes with respect to their relative heights. The raw trajectories are observed on sparse and unevenly spaced points and therefore a preliminary smoothing step is required. The original data were smoothed using a spline basis (see López-Pintado and Romo, 2009, for details about the data). Since the number of observations per tree is very irregular, in those curves with fewer observations the smoothness procedure is less effective as shown in Fig. 9(a) and (b). These curves will have low depth and could be treated as outliers. The deepest function based on S_H is represented in Fig. 9(a) and could be considered as a median curve representing in this case the typical profile of a Laricio tree. We also calculated the 10 deepest curves within the sample (see Fig. 9(b)) and illustrate that they are smooth and not affected by outlier curves. This shows that the median or trimmed mean could be more robust and representative within the sample than the sample mean.

In Fig. 10 the deepest Laricio trees profile based on the modified half-region depth are represented with thick lines. In the three real data examples the deepest curves obtained using the half-region depth are closer to each other than the

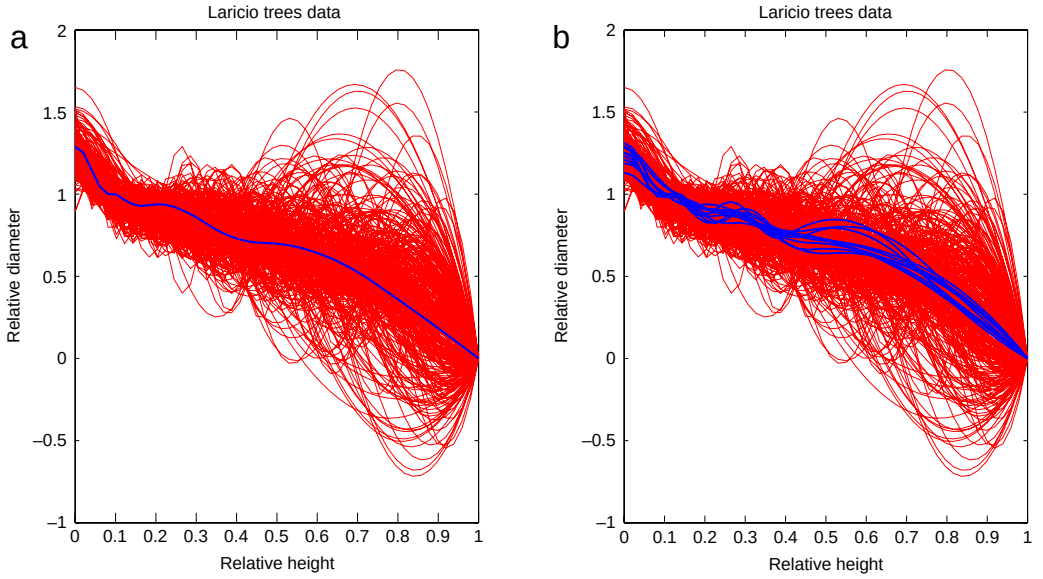

Fig. 9. Relative diameter versus relative height of a sample of 354 Laricio trees. The original raw curves were smoothed using a spline basis. (a) The deepest curve using the half-region depth is represented in a thick dark line and (b) the 10 deepest curves based on the half-region depth are also plotted using thick lines.

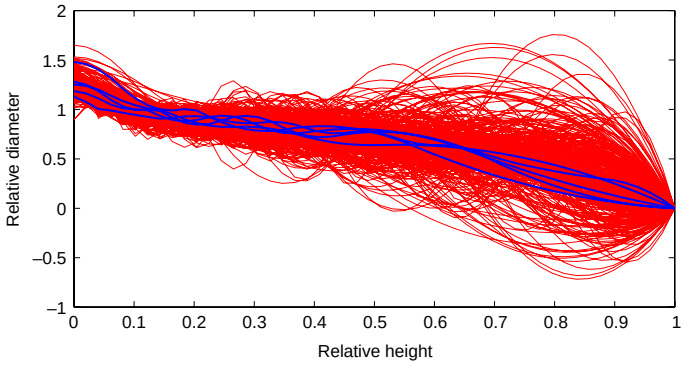

Fig. 10. Relative diameter versus relative height of a sample of 354 Laricio trees. We have represented the five deepest curves based on the modified half-region depth using thick lines.

ones obtained with the modified half-region depth. We recommend the use of the half-region depth when the curves are sufficiently smooth because of its simplicity, theoretical properties and robustness with respect to shape outliers.

7.2. High-dimensional data example

Most of the common multivariate notions of depth, such as simplicial depth (Liu, 1990) or half-space depth (Tukey, 1975) are not computationally feasible when the dimensionality of the data is greater than three. Moreover, with high-dimensional data the number of observations is usually smaller than the dimension of the data and traditional multivariate depths are not well defined in this situation. In contrast, the half-region depths can be defined and applied to high-dimensional data with no computational cost. An important high-dimensional data example of interest in bioinformatics is gene expression microarray data. Microarray experiments measure the level of expression of thousands of genes simultaneously. The data is high-dimensional since many variables (genes expression) are measured for each observation or tissue (samples). In Fig. 11 the level of expression of 50 genes in a sample of 47 individuals with acute lymphoblastic leukemia are represented using parallel coordinates. The genes are represented in the x axis and the expression of these genes in the y axis. Each curve corresponds to one individual from the sample. The deepest (or more representative) expression profile using the modified half-region depth are represented in dark black. Based on the center-outward order provided by the modified half-region depth we can define location estimates such as the median or trimmed mean. These robust statistics can be very useful in high-dimensional data since outliers are sometimes difficult to detect and can affect the results in many ways. The depth

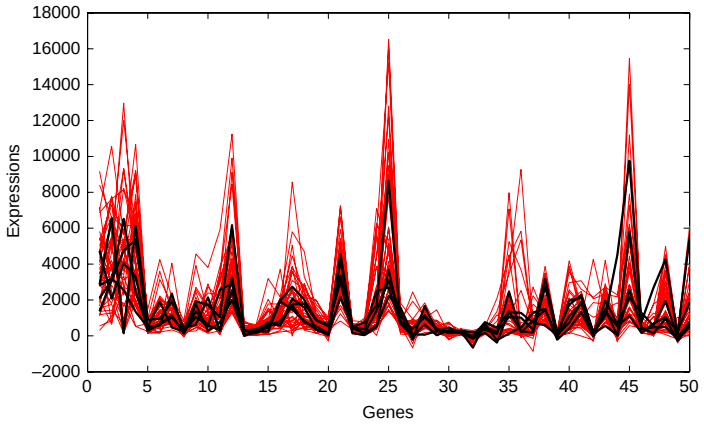

Fig. 11. Expression of 50 genes for 47 individuals with lymphoblastic leukemia. The deepest curves are represented in dark black.

based median or trimmed mean expression profile could be used as a building block for extending robust statistical methods to high-dimensional data (see López-Pintado et al., 2010).

8. Conclusions

We have introduced a simple new notion of depth for functional data based on the ideas of hypograph and epigraph of a curve. In contrast to other notions of depth, the half-region depth can be applied to high-dimensional data with little computational cost. Order statistics for functional data based on this depth, such as the median and the trimmed mean, can be defined. Several desirable properties of the half-region depth, in the finite- and infinite-dimensional cases, have been established such as continuity, consistency and uniform convergence of its sample version. A simulation study illustrates the performance in several contaminated models of the half-region depth compared to other proposed functional depths and location estimators. Half-region depth is competitive and in certain cases outperforms alternative methods. Finally, real data examples have been analyzed, and illustrate how the deepest curves are representative within the sample of curves.

Acknowledgements

This research was partially supported by Spanish Ministry of Education and Science grants BEC 2002-03769, SEJ2005-06454, SEJ2007-67734, SEJ2905 and ECO2008-05080.

References

Arcones, M., Chen, Z., Gine, E., 1994. Estimators related to U-processes with applications to multivariate medians: asymptotic normality. Annals of Statistics 22, 1460–1477.
Cuesta-Albertos, J.A., Nieto-Reyes, A., 2008. The random Tukey depth. Computational Statistics and Data Analysis 52 (11), 4979–4988.
Cuevas, A., Febrero, M., Fraiman, R., 2006. On the use of bootstrap for estimating functions with functional data. Computational Statistics and Data Analysis 51 (2), 1063–1074.
Cuevas, A., Febrero, M., Fraiman, R., 2007. Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics 22, 481–496.
Donoho, D., Gasko, M., 1992. Breakdown properties of location estimates based on half-space depth and projected outlyingness. Annals of Statistics 20, 1803–1827.
Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis. Springer, New York.
Fraiman, R., Muniz, G., 2001. Trimmed means for functional data. Test 10, 419–440.
González-Manteiga, W., Vieu, P., 2007. Statistics for functional data. Computational Statistics and Data Analysis 51, 4788–4792.
Inselberg, A., 1985. The plane parallel coordinates. Visual Computer 1, 69–91.
Liu, R.Y., 1990. On a notion of data depth based upon random simplices. Annals of Statistics 18, 405–414.
Liu, R.Y., Parelius, J.M., Singh, K., 1999. Multivariate analysis by data depth: descriptive statistics, graphics and inference. Annals of Statistics 27, 783–858.
López-Pintado, S., Jörnsten, R., 2007. Functional data analysis via extensions of the band depth. In: IMS Lecture Notes-Monograph Series Institute of Mathematical Statistics, vol. 54. pp. 103–120.
López-Pintado, S., Romo, J., 2009. On the concept of depth for functional data. Journal of the American Statistical Association 104 (486), 718–734.
López-Pintado, S., Romo, J., 2007. Depth-based inference for functional data. Computational Statistics and Data Analysis 51, 4957–4968.
López-Pintado, S., Romo, J., Torrente, A., 2010. Robust depth-based tools for the analysis of gene expression data. Biostatistics 11 (2), 254–264.
Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proceedings of National Academy of Science of India 12, 49–55.
Oja, H., 1983. Descriptive statistics for multivariate distributions. Statistics and Probability Letters 1, 327–332.
Pollard, D., 1984. Convergence of Stochastic Processes. Springer Verlag, New York.
Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, second ed. Springer Verlag, New York.
Tukey, J., 1975. Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians. Vancouver, pp. 523–531.
Van Der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press.
Wegman, E., 1990. Hyperdimensional data analysis using parallel coordinates. Journal of the American Statistical Association 85, 664–675.
Wood, A.T.A., Chan, G., 1994. Simulation of stationary gaussian processes in $C[0, 1]$. Journal of Computational and Graphical Statistics 3, 409–432.
Zuo, Y., 2003. Projected based depth functions and associated medians. Annals of Statistics 31, 1460–1490.
Zuo, Y., Serfling, R., 2000. General notions of statistical depth functions. Annals of Statistics 28, 461–482.