UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 05-16
Statistics and Econometrics Series 03
Abril 2005

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

A HALF-GRAPH DEPTH FOR FUNCTIONAL DATA

Sara López-Pintado y Juan Romo*

Abstract

A recent and highly attractive area of research in statistics is the analysis of functional data. In this paper a new definition of depth for functional observations is introduced based on the notion of "half-graph" of a curve. It has computational advantages with respect to other concepts of depth previously proposed. The half-graph depth provides a natural criterion to measure the centrality of a function within a sample of curves. Based on this depth a sample of curves can be ordered from the center outward and *L*-statistics are defined. The properties of the half-graph depth, such as the 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.

Key Words: Functional data, data depth, order statistics.

*López-Pintado, Departamento de Estadística, Universidad Carlos III de Madrid, e-mail: saral@est-econ.uc3m.es; Romo, Departamento de Estadística, Universidad Carlos III de Madrid, e-mail: juan.romo@uc3m.es.

A half-graph depth for functional data

Sara López-Pintado and Juan Romo
Universidad Carlos III de Madrid

April 7, 2005

Abstract

A recent and highly attractive area of research in statistics is the analysis of functional data. In this paper a new definition of depth for functional observations is introduced based on the notion of "half-graph" of a curve. It has computational advantages with respect to other concepts of depth previously proposed. The half-graph depth provides a natural criterion to measure the centrality of a function within a sample of curves. Based on this depth a sample of curves can be ordered from the center outward and L-statistics are defined. The properties of the half-graph depth, such as the 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.

Key words: Functional data, data depth, order statistics.

1 Introduction

The analysis of functional data is one of the topics that, within the field of statistics, is receiving a steadly increasing attention in recent years (see for example Ramsay and Silverman (1997)). A fundamental task in functional data analyisis is to provide a natural ordering within a sample of curves, thus making possible to define ranks and L-statisitcs. In this paper we introduce a new definition of depth for functional observations based on the concepts of hypergraph and hypograph of a curve. This functional depth provides a criterion to order the sample of curves from center outward. The notion of statistical depth was first analyzed for multivariate observations and different definitions of depth have been studied in the literature: Mahalanobis (1936), Tukey (1975), Liu (1990), Oja (1983), Singh (1991), Donoho and Gasko (1992), Zuo and Serfling (2000) and Zuo (2003), among others. These multivariate depths are not adequate for high-dimensional data, therefore, their applicability is restricted to vector observations with low dimensions. An advantage of the graph-based depth is that it is computationally not very intensive and can be easily adapted to high-dimensional data. The paper is organized as follows. In the next (second) section we define the new concept of functional depth, referred as half-graph depth S_H. In section three we analyze the finite-dimensional version of S_H and some of its properties, such as the consistency and the uniform convergence, are established. We extend these results to the infinite-dimensional case in section four. Section five deals with a generalized version of S_H that is more convenient for non-smooth functional data. Throughout section six simulated curves are considered to show the performance of these functional depths and, finally, in the last section, real data examples are analyzed.

2 Half-graph depth

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

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

Define the hypograph *(hg)* and the hypergraph *(Hg)* of a function x in $C(I)$ as

$$\begin{aligned} hg(x) &= \{(t,y) \in I \times \mathbb{R} : \ y \leq x(t) \ \}, \\ Hg(x) &= \{(t,y) \in I \times \mathbb{R} : y \geq x(t) \ \}. \end{aligned}$$

Figures 1 and 2 give, respectively, the hypograph and hypergraph of a curve x.

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

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

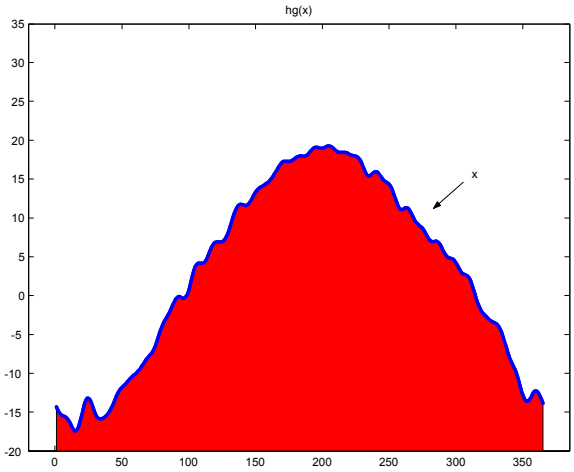

Figure 1: *Hypograph of the function x.*

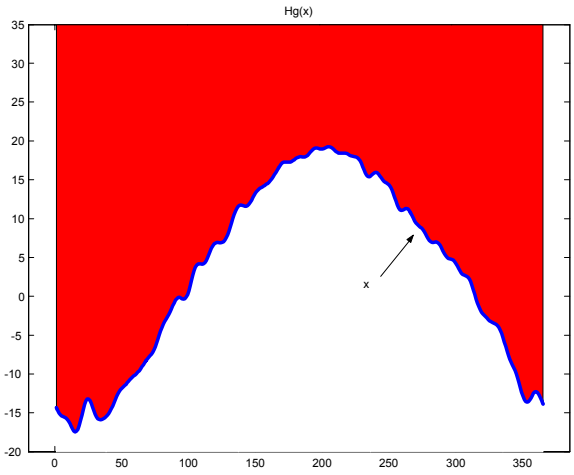

Figure 2: *Hypergraph of the function x.*

where

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

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

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

Hence, the half-graph 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 hypergraph of x.

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

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

where

$$\begin{aligned} G_1(x) &= P\left(G(X) \subset hg(x)\right) \\ &= P\left(X(t) \leq x(t),\ t \in I\right), \end{aligned}$$

and

$$\begin{aligned} G_2(x) &= P\left(G(X) \subset Hg(x)\right) \\ &= P\left(X(t) \geq x(t),\ t \in I\right). \end{aligned}$$

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

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

where

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

and

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

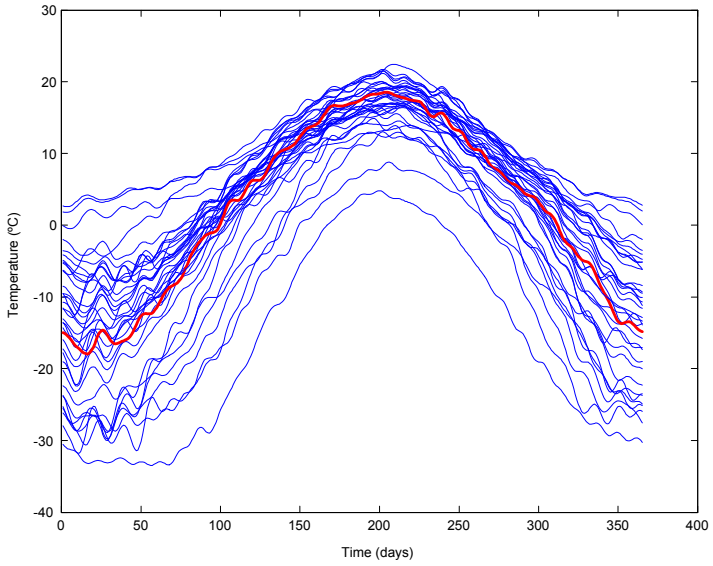

Figure 3: *Daily temperatures during one year in 35 weather stations in Canada; the curve which maximizes the depth $S_{n,H}$ is represented in red.*

The sample version of this second way of defining the half-graph depth is obtained 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-graph depth,

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

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, $x_2, ..., x_n$ are ordered according to decreasing values of $S_{n,H}(x_i)$ we obtain order statistics $x_{(1)}$, $x_{(2)}, ..., x_{(n)}$, where $x_{(1)}$ denotes the deepest observation and $x_{(n)}$ the less deepest one. Figure 3 shows a real data example that consists of the daily temperatures during one year in thirty five different weather stations in Canada. The curve represented in red color is the deepest one.

3 Finite-dimensional version

The concepts of hypograph and hypergraph introduced in the previous section can be easily adapted to finite-dimensional data. Consider each point in $\mathbb{R}^d$ as a real function defined in the set of indexes $\{1, ..., d\}$, the hypograph and hypergraph of a point $x = (x(1), x(2)..., x(d))$ can be expressed respectively as

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

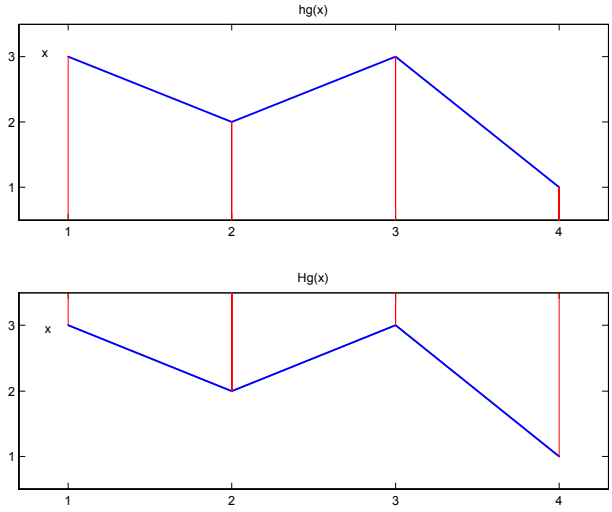

Figure 4: *Hypograph and hypergraph of a point in* $\mathbb{R}^4$ *using parallel coordinates.*

and

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

Figure 4 gives the hypograph and the hypergraph of a point $x = (3,\ 2,\ 3,\ 1) \in \mathbb{R}^4$ using parallel coordinates (Inselberg, 1985). In Figure 5 two points belonging to the $hg(x)$ and $Hg(x)$ are represented (using green color). An alternative interpretation can be obtained using the cartesian representation of the points in $\mathbb{R}^d$ (with $d \leq 3$). Figure 6 shows the hypograph and hypergraph of a point x in $\mathbb{R}^2$ using its representation in cartesian coordinates.

Let X be a d-dimensional random variable with distribution function F. $X \leq x$ and $X \geq x$ are the abreviations for $\{X(k) \leq x(k),\ k = 1, ..., d\}$ and $\{X(k) \geq x(k),\ k = 1, ..., d\}$, respectively. If we particularize the half-graph 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 $x_1, ..., x_n$ be a random sample from the variable X, the sample version of the half-graph depth is

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

Figure 7 shows 50 points simulated from a normal bivariate distribution and it ilustrates the way of computing the half-graph depth $S_{n,H}$ of a point from the sample represented in red. The

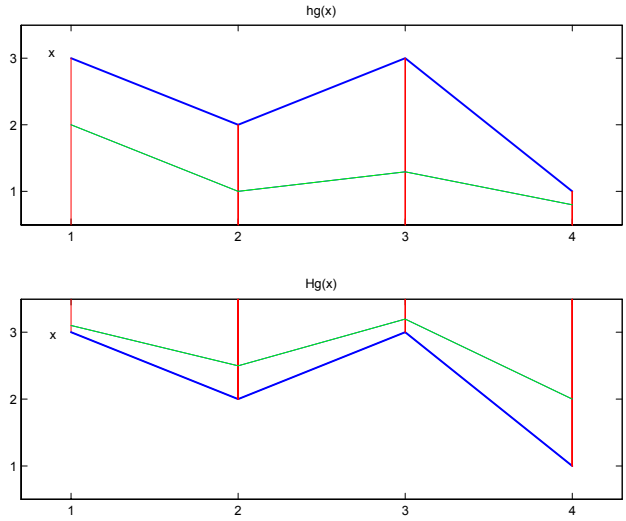

Figure 5: *The points* (2, 1, 1.3, 0.8) *and* (3.1, 2.5, 3.2, 2) *belonging to the hypograph and hypergraph of* x*, appear respectively in green color.*

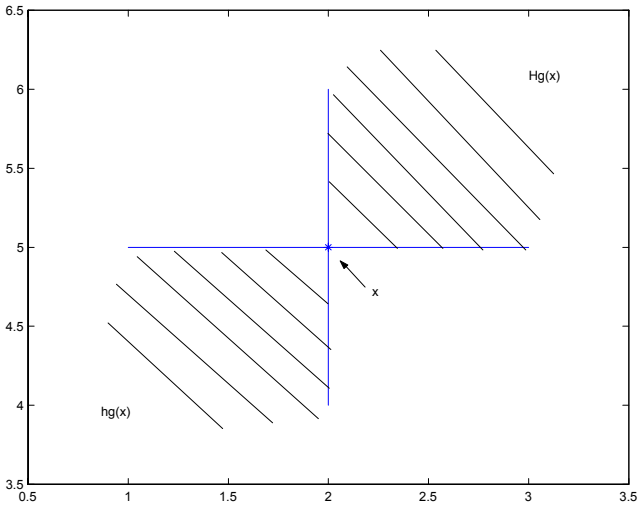

Figure 6: *Finite-dimenisonal version of the hypergraph and the hypograph of a point* $x \in \mathbb{R}^2$.

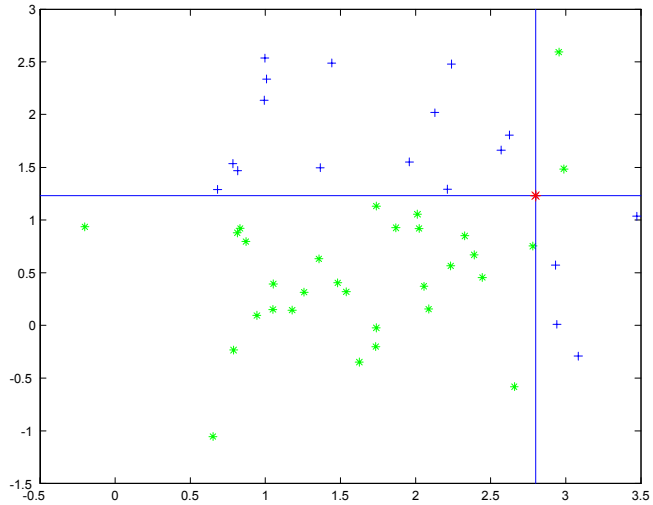

Figure 7: *Half-graph depth of a point from a sample of 50 points from a normal distribution.*

proportion of points from the sample in the upper right cuadrangle (hypergraph) is 3/50 and in the lower left cuadrangle (hypograph) is 30/50; hence, the half-graph depth of the point in red is 3/50. In Figure 8 the deepest point from the same sample appears in red. We also illustrate the way of computing the depth of this point. The proportion of data in the upper right cuadrangle is 12/50 and in the lower left cuadrangle is 17/50; therefore, the half-graph depth of the deepest point is 12/50.

The half-graph depth is invariant with respect to traslations 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 2 *For $d = 1$ the half-graph depth $S_H(x)$ can be expressed as*

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

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

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

Proposition 3 *(Vanishing at infinity) Let $x \in \mathbb{R}^d$,*

$$\sup_{\|x\| \geq M} S_H(x) \longrightarrow 0, \qquad \textit{when } M \to \infty.$$

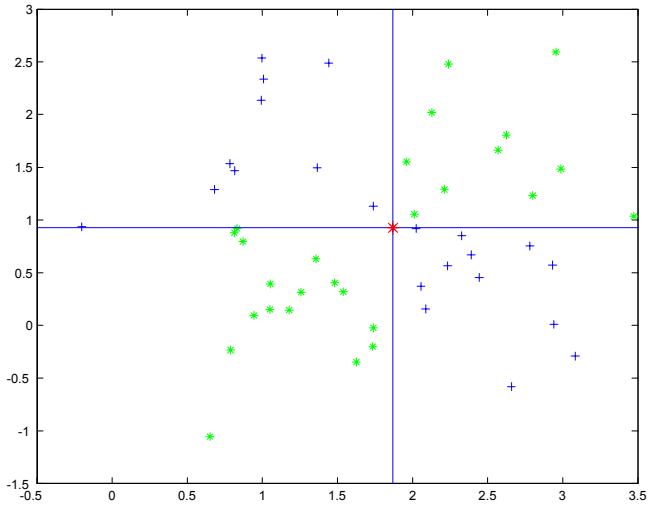

Figure 8: *Deepest point in red from a sample of 50 normal observations.*

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

Note that the previous proposition implies that

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

Proposition 4 *$S_H(\cdot)$ is an upper semicontinuous function. Moreover, if F is absolutely continuous then $S_H(\cdot)$ is continuous.*

The proofs of Propositions 3 and 4 are postponed to next section, since they are particular cases of the same properties in the functional case. In the next proposition we prove the uniform convergence of $S_{n,H}$ to its population version.

Proposition 5 *$S_{n,H}$ is uniformly consistent:*

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

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

$$\tau_n \xrightarrow{a.s.} \tau, \qquad \textit{when } n \to \infty.$$

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,

$$\sup_{x\in\mathbb{R}^d} |\min\{F_{X_n}(x), F_{Y_n}(y)\} - \min\{F_X(x), F_Y(y)\}|$$

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

The second part of the theorem is proven using arguments similar to the ones proposed by Arcones *et al.* (1994) to show the consitency of the simplicial median. By Proposition 4, S_H is an upper semicontinuous function, then, $\limsup_{n\to\infty} S_H(y_n) \leq S_H(y)$, 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 $\varepsilon > 0$, $S_H(h) - \sup_{|x-h|\geq\varepsilon} S_H(x) > 0$. Hence, for the following argument consider $\delta = S_H(h) - \sup_{|x-h|\geq\varepsilon} S_H(x) > 0$.

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

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

Recall that

$$P\left\{\sup_{n\geq l} |\tau_n - \tau| > \varepsilon\right\} \leq P\left\{\sup_{n\geq l}(S_H(\tau) - S_H(\tau_n)) \geq \delta\right\}$$

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

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

4 Properties of the functional depth

Here, we extend some of the properties established in the previous section to the functional version of the half-graph depth S_H. Let $x_1, ..., x_n$ 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.,

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

The depth verifies 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-graph depth of a function converges to zero when its norm tends to infinity.

Proposition 6 *The depths S_H and $S_{n,H}$ verify that*

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

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

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,

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

And also,

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

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 $\|x\|_\infty = \sup x(t)$,

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

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

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

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

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

It is easy to see that the following inequality holds,

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

and it implies that

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

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

Proposition 7 *$S_H(\cdot)$ 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 $y_n \xrightarrow{\|\cdot\|_\infty} y$.

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

To establish the continuity of the functional $S_H(\cdot)$ in $C(I)$ with respect to the supremum norm 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

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

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

$$\begin{aligned} P\left(G(X) \subset hg(x_n) \cap G(X) \subsetneq hg(x)\right) &\underset{n\to\infty}{\longrightarrow} 0 \\ P\left(G(X) \subsetneq hg(x_n) \cap G(X) \subset hg(x)\right) &\underset{n\to\infty}{\longrightarrow} 0. \end{aligned} \tag{4}$$

Hence, G_1 is a continuous function. ■

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

Theorem 8 *$S_{n,H}$ is strongly consistent,*

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

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

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

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

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

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-graph depth can be expressed as a transformation of two empirical processes. We present first 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

$$Pf = \int f(X)dP.$$

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

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

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

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

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

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

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

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

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-graph depth for classes of functions $\mathcal{F}_1$ and $\mathcal{F}_2$ with finite bracketing number.

Theorem 9 *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 provide now some examples of families of functions $\mathcal{F}$ that verify the condition of finite bracketing number. In addition, a continuity condition for the probability distirbution 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 10 *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 $t, s \in I$,*

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

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

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

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

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

Since the second term converges almost surely to zero when M tends to infinity by Proposition 6, given M sufficiently large, we just need to prove that

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

We have that

$$\sup_{x \in E_M} |S_{n,H}(x) - S_H(x)| \leq$$

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

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

$$\sup_{x \in E_M} |S_{n,H}(x) - S_H(x)| \leq$$

$$\leq \sup_{f \in \mathcal{F}_1^M} |P_n f - P f| + \sup_{f \in \mathcal{F}_2^M} |P_n f - P f|.$$

Therefore, it is sufficient to establish that the families $\mathcal{F}_1^M$ and $\mathcal{F}_2^M$ have finite bracketing number ($\mathcal{N}_{[\,]}\left(\varepsilon, \mathcal{F}_1^M, L_1(P)\right) < \infty$ and $\mathcal{N}_{[\,]}\left(\varepsilon, \mathcal{F}_2^M, L_1(P)\right) < \infty$). 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 $z_1, ..., z_p$ that determine brackets covering the family $\mathcal{F}_1^M$ and verifying

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

Since the condition C1 is verified, 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 $[0, \delta), [\delta, 2\delta), [2\delta, 3\delta), ..., [([1/\delta] - 1)\delta, 1)$, and taking values in the secuence: $-[M/\nu]\nu, ... - \nu, 0, \nu, 2\nu, ..., [M/\nu]\nu$. The total number p of possible functions that can be constructed like this is finite and we denote them as $z_1, ..., z_p$. Define the following set of indicator functions $f_z : C \to \{0, 1\}$,

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

This set of functions allows to construct $\varepsilon-$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

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

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

$$\begin{aligned} I_{\{z_i \leq X\}} &\leq I_{\{x \leq X\}} \leq I_{\{z_j \leq X\}} \\ f_{z_i} &\leq f_x \leq f_{z_j} \end{aligned}$$

and

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

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 12 *Let P be a distribución verifying 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*

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

Proof. We have to show that

$$P\left(\sup_{n \geq l} \|\tau_n - \tau\|_\infty \geq \varepsilon\right) \underset{l \to \infty}{\longrightarrow} 0.$$

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

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

Then the proof is analogous to the one in Proposition 5. ■

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 verifies Theorem 11. Hence, the sample half-graph depth $S_{n,H}$ converges uniformly to S_H over the set $Lip_{\alpha,A}(I)$.

4.1 A generalized half-graph depth

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

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

where λ stands for the Lebesgue's 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 generalized half-graph depth at x is:

$$GS_H(x) = \min\left\{SL(x), IL(x)\right\}.$$

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

$$GS_{n,H}(x) = \min\left\{SL_n(x), IL_n(x)\right\},$$

where

$$\begin{aligned} SL_n(x) &= \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\left\{t \in I : x(t) \leq x_i(t)\right\}, \\ IL_n(x) &= \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\left\{t \in I : x(t) \geq x_i(t)\right\}. \end{aligned}$$

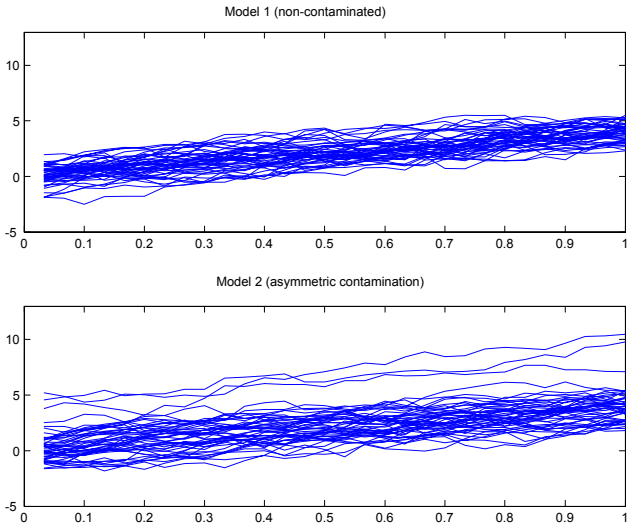

Figure 9: *Curves generated with the non-contaminated model and the asymmetric contaminated model.*

5 Simulation results

In this section, we report the results of a simulation study where the robustness of the half-graph and generalized half-graph depths is analyzed. We have simulated curves from different contaminated models and compared the trimmed means estimates (based on S_H and GS_H) with those obtained using the mean. The models considered are: an elementary non-contaminated model M1, and four contaminated models denoted as: M2, M3, M4, M5. Some of these models were analyzed by Fraiman and Muniz (2001). The elementary model M1 consists of p curves verifying

$$X_i(t) = f(t) + e_i(t), \qquad 1 \leq i \leq p$$

where $e_i(t)$ is a Gaussian stochastic process with zero mean and covariance function

$$E(e_i(t)e_i(s)) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{5|t-s|}$$

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

The asymmetric total contamination model (M2) is defined by

$$Y_i(t) = X_i(t) + \epsilon_i M \qquad 1 \leq i \leq p,$$

where ϵ_i takes values 1 with probability q and 0 with probability $1 - q$. The constant M is the contamination size. In Figure 9 we represent curves simulated using the non-contaminated and asymmetric contaminated models.

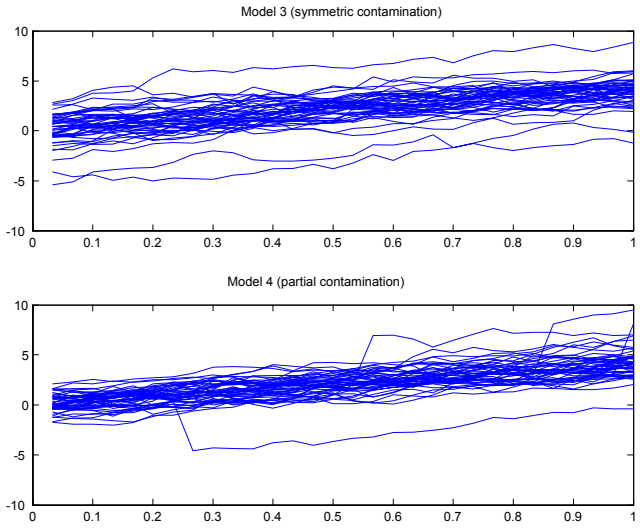

Figure 10: *Curves from the symmetric and partially contaminated models.*

A model of symmetric contamination (M3) can be obtained in the following way:

$$Y_i(t) = X_i(t) + \epsilon_i \sigma_i M \qquad 1 \leq i \leq p,$$

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

A partially contaminated model (M4) can be expressed as follows:

$$\begin{aligned} Y_i(t) &= X_i(t) + \epsilon_i \sigma_i M \qquad \text{for} \quad t \geq T_i \quad 1 \leq i \leq p, \quad \text{and} \\ Y_i(t) &= X_i(t) \qquad\qquad\qquad\quad \text{for } t < T_i \end{aligned}$$

where ϵ_i, M and σ_i are defined as in model 3 and T_i is a random number generated from a uniform distribution on $(0, 1)$. Figure 10 shows curves generated from the symmetric and partially contaminated models.

Finally, the fifth model considered here is a pick contamination model (M5) expressed as:

$$\begin{aligned} &Y_i(t) = X_i(t) + \epsilon_i \sigma_i M, \qquad \text{for} \quad T_i \leq t \leq T_i + l, \quad 1 \leq i \leq p, \quad \text{and} \\ &Y_i(t) = X_i(t), \qquad\qquad\qquad\quad \text{for } t \notin [T_i, T_i + l], \end{aligned}$$

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 Figure 11 we represent curves generated from the picks contaminated model.

For each model we have considered $N = 500$ replications for $p = 50$ curves and contamination fraction $q = 0.1$. In addition we use two contamination constants $M = 5$, $M = 25$ and trimming level equal to $\alpha = 0.2$.

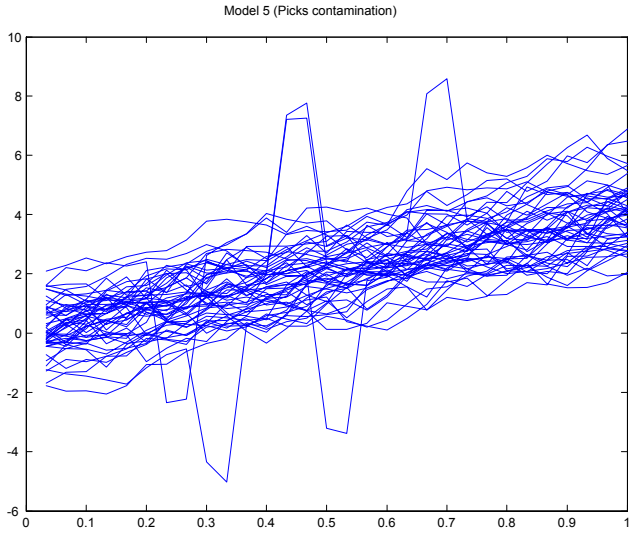

Figure 11: *Curves generated from the picks contaminated model.*

We have calculated the mean and the trimmed mean estimates in every model,

$$\widehat{m}_n(t) = \frac{\sum_{i=1}^{p} X_i(t)}{p}$$

and

$$\widehat{m}_{n,\alpha}(t) = \frac{\sum_{i=1}^{p-[p\alpha]} X_{(i)}(t)}{p - [p\alpha]}$$

where $\alpha = 0.2$, and $[p\alpha]$ is the integer part of $p\alpha$. For each of the N replications the integrated error is calculated and is evaluated at $I = 30$ equally spaced points in $[0, 1]$,

$$EI_m(j) = \frac{1}{I}\sum_{k=1}^{I} \left[\widehat{g}_n(k/I) - f\left(k/I\right)\right]^2$$

where $\widehat{g}_n$ is either $\widehat{m}_n$ or $\widehat{m}_{n,\alpha}$. In tables 1 and 2 we represent the mean integrated error for each estimate, $\widehat{m}_n$ and $\widehat{m}_{n,\alpha}$ defined as:

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

and its standard deviation

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

We estimate the mean integrated error using the sample mean (E_m), the $\alpha-$trimmed mean based on S_H ($E^{\alpha}_{S_H}$) and the $\alpha-$trimmed mean based on GS_H ($E^{\alpha}_{GS_H}$).

Estim.	**M1**	**M2**	**M3**	**M4**	**M5**
E_m	0.0104	0.3106	0.0614	0.0342	0.0916
(S_m)	(0.0076)	(0.2563)	(0.011)	(0.0367)	(0.0477)
$E^{0.2}_{S_H}$	0.0133	0.3293	0.0296	0.0416	0.1116
$\left(S^{0.2}_{S_H}\right)$	(0.0101)	(0.3036)	(0.0502)	(0.0452)	(0.0648)
$E^{0.2}_{GS_H}$	0.0134	0.0190	0.0145	0.0225	0.1198
$(S^{0.2}_{GS_H})$	(0.0106)	(0.0306)	(0.0107)	(0.0201)	(0.0649)

Table 1: *N=500, p=50, q=0.1 and M=5.*

Estim.	**M1**	**M2**	**M3**	**M4**	**M5**
E_m	0.0098	7.3973	1.2307	0.5947	0.8566
(S_m)	(0.0067)	(6.0935)	(1.7753)	(0.7722)	(0.0577)
$E^{0.2}_{S_H}$	0.126	6.8191	0.3566	0.7694	0.8519
$\left(S^{0.2}_{S_H}\right)$	(0.0082)	(6.7457)	(1.7753)	(0.9506)	(0.0549)
$E^{0.2}_{GS_H}$	0.0126	0.1037	0.0247	0.3173	1.3122
$(S^{0.2}_{GS_H})$	(0.0086)	(0.6365)	(0.1514)	(0.4883)	(0.0649)

Table 2: *N=500, p=50, q=0.1 and M=25.*

Table 1 provides the results of the simulation with $N = 500$, $p = 50$, $q = 0.1$ and $M = 5$. The mean integrated errors in models M2, M3 and M4 are minimized using the generalized half-graph depth. In the remaining models the mean gives the best results. In table 2 the contamination constant is now $M = 25$. The minimum mean integrated errors in models M2, M3 and M4 are again obtained with GS_H and in model M5 the mean integrated error is minimized with S_H.

6 Real data examples

Herein, we describe some real data examples to illustrate the performance of the half-graph depth and the generalized half-graph depth. The first example, introduced by Ramsay and

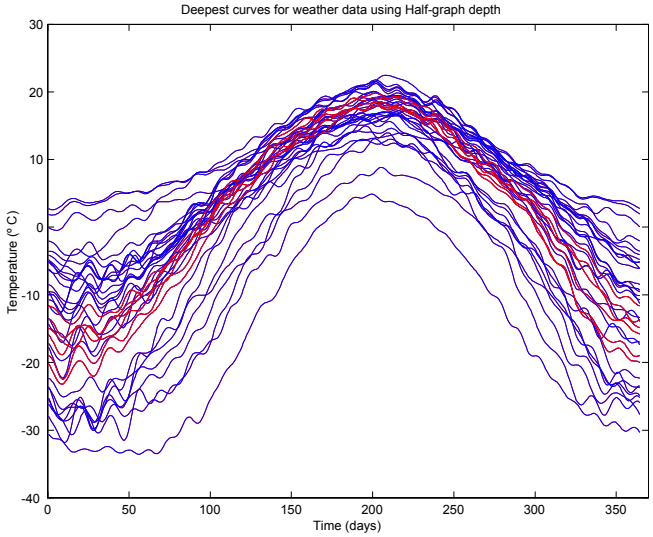

Figure 12: *Temperatures in different weather stations in Canada during one year. The curves in red are the deepest curves using the half-graph depth.*

Silverman (1997), consists on the temperature recorded along a year in thirty nine different weather stations in Canada (Figure 12). These curves have been smoothed using a Fourier basis.

Figure 13 shows another examble based on the one year rainfall in the same Canadian weather stations.

The deepest curves (represented in red) have two shapes, thus suggesting that rain curves follow two major patterns. Although with oscillations, on average some of the curves are increasing the first two hundred days of the year and after reaching a maximum they decrease; in contrast, other curves decrease during the first two hundred days and after a minimum they increase.

A third real data example consists on the growth curves of a sample of girls (Figure 14). The deepest curves obtained using the half-graph depth are represented in red.

Finally, we have applied the half-graph depth to analyze how the relative diameter of a sample of Laricio trees changes with respect to their relative heights. The original data were smoothed using a spline basis. Since the number of observations per tree is very irregular, in those curves with fewer observations the smoothness procedure is less effective. The deepest observation and the ten deepest curves are represented in figures 15 and 16, respectively.

The results of the analysis of all these real data examples using the generalized half-graph depths are shown in figures 17 to 20.

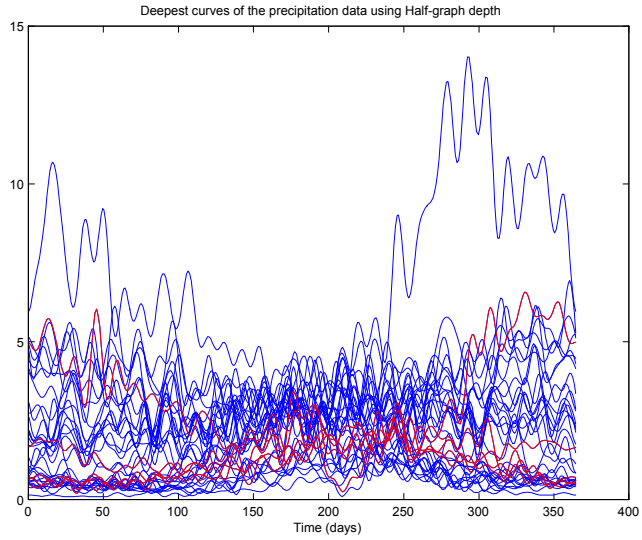

Figure 13: *Representation of the rain fallen in thirty nine weather stations in Canada during one year. The six deepest observations using S_H are represented in red.*

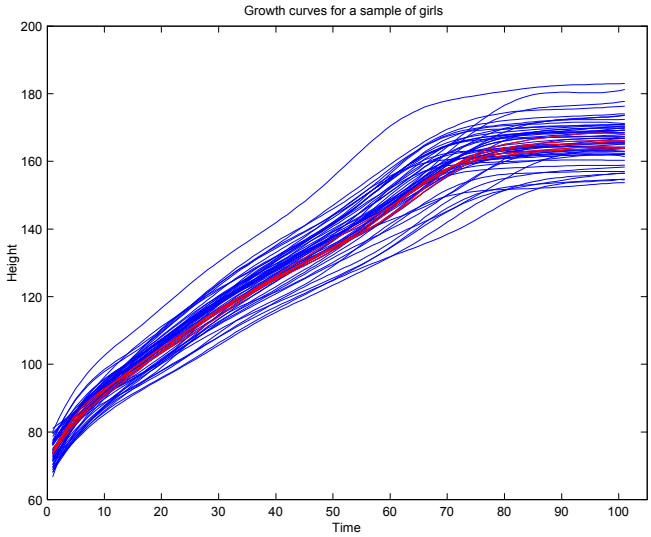

Figure 14: *Heights in cm of fifty four girls during their first ten years of life. The original curves were smoothed using a spline basis. In red we have represented the deepest curves based on S_H.*

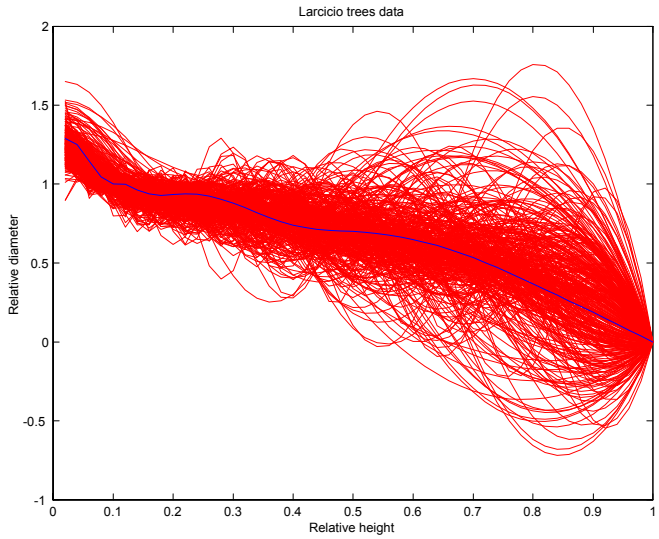

Figure 15: *Relative diameter versus relative height of a sample of three hundred and fifty four Laricio trees. The original data was smoothed using a spline basis. The curve in blue is the deepest curve.*

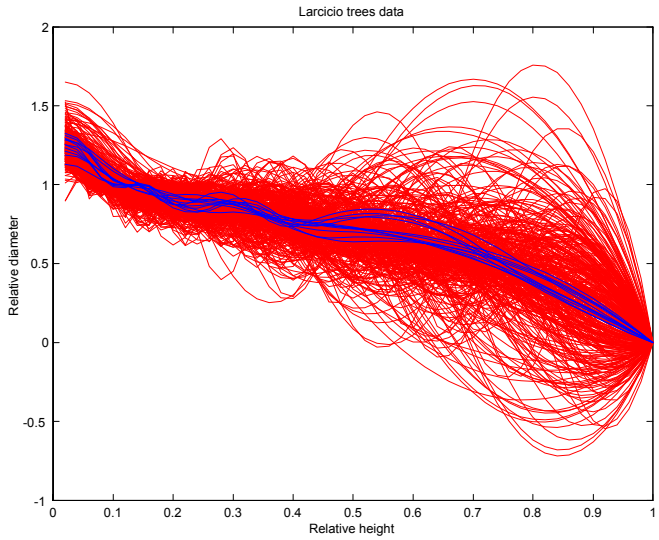

Figure 16: *Representation of the ten deepest curves from the Laricio trees example.*

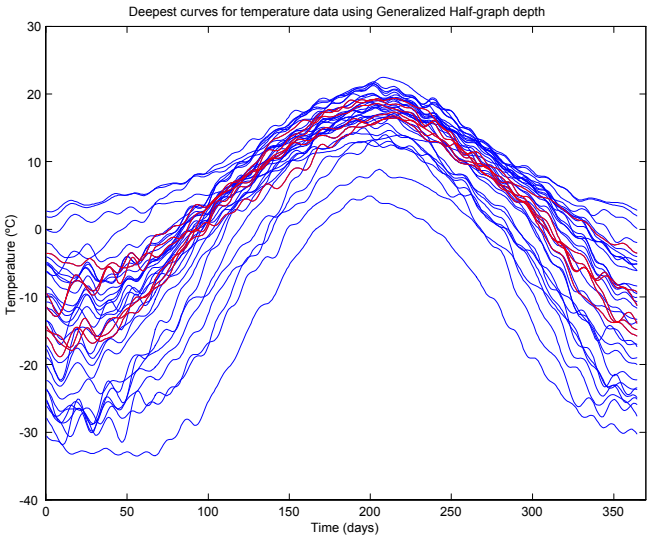

Figure 17: *Temperatures in different weather stations in Canada during one year. The curves in red are the deepest curves using the generalized half-graph depth.*

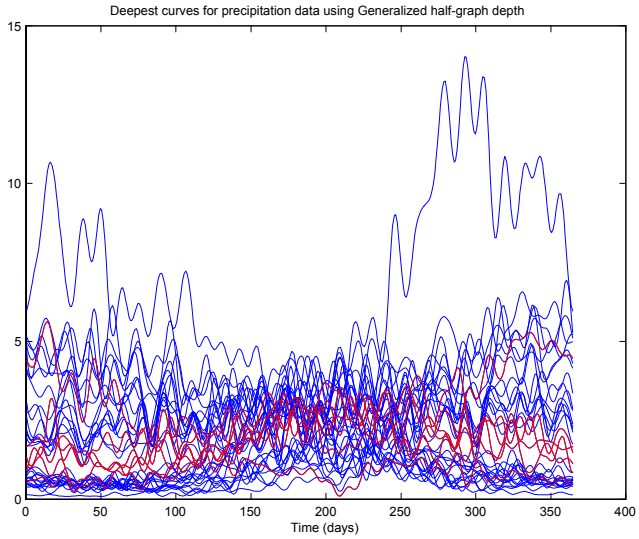

Figure 18: *Rain fallen in thirty nine weather stations in Canada during one year. The six deepest observations are represented in red.*

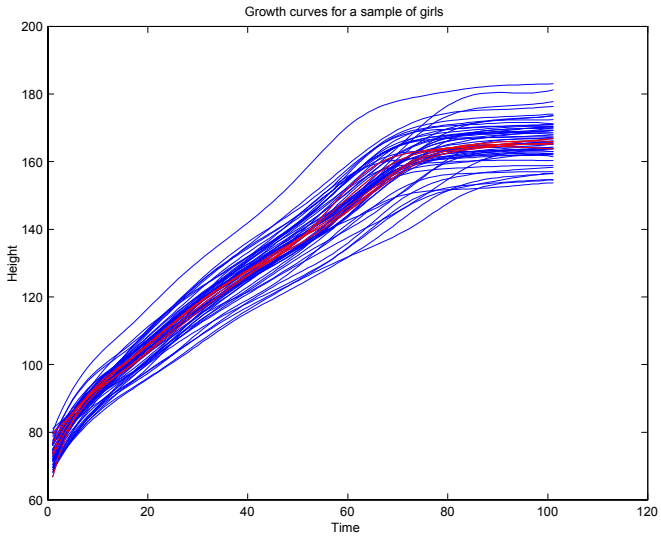

Figure 19: *Heights in cm of fifty four girls during their first ten years of life. The original curves were smoothed using a spline basis. In red we have represented the deepest curves based on* GS_H.

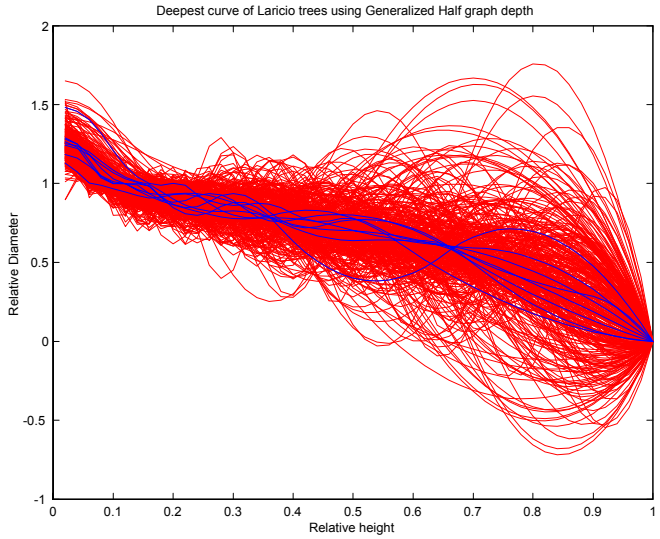

Figure 20: *Representation of the eight deepest curves based on* GS_H *from the Laricio trees example.*

References

[1] Arcones, M., Chen, Z. and Gine E. (1994). Estimators related to U-processes with applications to multivariate medians: asymptotic normality. *Ann. Statist.* **22**, 1460-1477.

[2] Donoho, D. and Gasko, M. (1992). Breakdown properties of location estimates based on half-space depth and projected outlyingness. *Ann. Statist.* **20**, 1803-1827.

[3] Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test* **10**, 419-440.

[4] Inselberg, A. (1985). The plane with parallel coordinates. Invited paper. *Visual Computer* **1**, 69-91.

[5] Liu, R. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18**, 405-414.

[6] Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proc. Nat. Acad. Sci. India* **12**, 49-55.

[7] Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag.

[8] Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1**, 327-332.

[9] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis.* Springer-Verlag.

[10] Zuo, Y. and Serfling, R.J. (2000). General notions of statistical depth function. *Ann. Statist.* **28**, 461-482.

[11] Singh, K. (1991). Majority depth. Unpublished manuscript.

[12] Tukey, J. (1975). Mathematics and picturing data. In *Proceedings of the 1975 International Congress of Mathematics* **2**, 523-531.

[13] Van Der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press.

[14] Zuo, Y. (2003). Projected based depth functions and associated medians. *Ann. Statist.* **31**, 1460-1490.