UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 06-31
Statistics and Econometrics Series 13
May, 2006

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

DEPTH-BASED INFERENCE FOR FUNCTIONAL DATA

Sara López-Pintado and Juan Romo*

Abstract

We propose robust inference tools for functional data based on the notion of depth for curves. We extend the ideas of trimmed regions, contours and central regions to functions and study their structural properties and asymptotic behavior. Next, we introduce a scale curve to describe dispersion in a sample of functions. The computational burden of these techniques is not heavy and so they are also adequate to analyze high-dimensional data. All these inferential methods are applied to different real data sets.

Key words: Functional data; Data depth; Trimmed regions; Scale curve.

*López-Pintado, Rutgers University, e-mail: saral@stat.rutgers.edu; Romo, Departamento de Estadística, Universidad Carlos III de Madrid, e-mail: juan.romo@uc3m.es.

Depth-based inference for functional data

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

[a] *Department of Statistics, Universidad Carlos III de Madrid*

[b] *Department of Statistics, Universidad Carlos III de Madrid*

Abstract

We propose robust inference tools for functional data based on the notion of depth for curves. We extend the ideas of trimmed regions, contours and central regions to functions and study their structural properties and asymptotic behavior. Next, we introduce a scale curve to describe dispersion in a sample of functions. The computational burden of these techniques is not heavy and so they are also adequate to analyze high-dimensional data. All these inferential methods are applied to different real data sets.

Key words: Functional data, Data depth, Trimmed regions, Scale curve

1 Introduction

The wide availability of functional data makes necessary to have robust inference tools for curves. The idea of statistical depth has been recently extended to functional observations. Fraiman and Muniz (2001) proposed a definition of depth as the integral of univariate depths. López-Pintado and Romo (2006) have introduced and studied the band depth, which is a notion of functional depth based on the curves graphs. The methods we construct and apply below can be implemented with any concept of depth for functional observations, but in this paper we will focus on the notions presented in the second of these papers.

Depth was introduced to generalize to multivariate observations ideas as median, order statistics or ranks, that are well established for univariate data.

⋆ Partially supported by grant SEJ2005-06454 from the Spanish Ministry of Education and Science

* Corresponding author: Tel: +916249805; fax: +916249849
Email address: juan.romo@uc3m.es (Juan Romo).

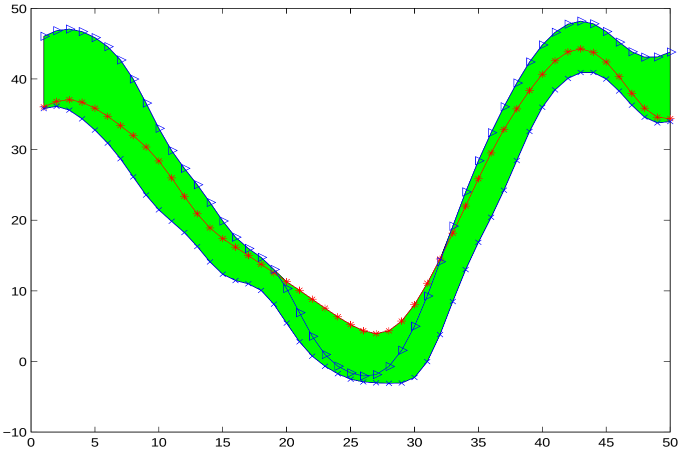

Fig. 1. *Band defined by three curves.*

Among others, some definitions have been provided by Mahalanobis (1936), Tukey (1975), Oja (1983), Liu (1990), Singh (1991), Vardi and Zhang (2001) and Zuo (2003). These notions of depth allow to construct a robust nonparametric inference for finite-dimensional observations (see Liu et al., 1999). In this paper, we extend these ideas to a functional context.

We recall next the band depth definitions and properties that we need throughout the paper. Let $C\,(I)$ be the set of continuous functions defined on the compact interval I in $\mathbb{R}$. Let $x_1(t), ..., x_n(t)$ be a collection of observations belonging to $C\,(I)$. The graph of a function x is the subset of $\mathbb{R}^2$ given by

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

The band in $\mathbb{R}^2$ defined by the curves $x_{i_1}, ..., x_{i_k}$ is

$$\begin{aligned} B(x_{i_1}, x_{i_2}, ..., x_{i_k}) &= \Big\{(t, y) :\ t \in I,\ \min_{r=1,\ldots,k} x_{i_r}(t) \le y \le \max_{r=1,\ldots,k} x_{i_r}(t)\Big\} \\ &= \Big\{(t, y) :\ t \in I,\ y = \alpha_t \min_{r=1,\ldots,k} x_{i_r}(t) + (1-\alpha_t) \max_{r=1,\ldots,k} x_{i_r}(t), \\ &\quad \text{for some } \alpha_t \in [0,1]\Big\}\,. \end{aligned}$$

Figure 1 shows the band defined by three curves that is the region in the plane enclosed by all of them.

For any function x in $C(I)$,

$$S_n^{(j}(x) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} I\{G(x) \subset B(x_{i_1}, x_{i_2}, ..., x_{i_j})\},\ j \ge 2, \qquad (1)$$

expresses the proportion of bands $B(x_{i_1}, x_{i_2}, ..., x_{i_j})$ given by j different curves $x_{i_1}, x_{i_2}, ..., x_{i_j}$ containing the graph of x. ($I\{A\}$ is one if A is true and zero otherwise). Given the sample $x_1, ..., x_n$, the *band depth* of x is

$$S_{n,J}(x) = \sum_{j=2}^{J} S_n^{(j}(x), \;\; J \geq 2. \tag{2}$$

Let $X_1, X_2, ..., X_n$ be independent copies of a stochastic process X with probability distribution P generating the sample $x_1, ..., x_n$; the population versions of these depth indexes are

$$S^{(j}(x, P) = S^{(j}(x) = P\left(G(x) \subset B(X_1, X_2, ..., X_j)\right)$$

and

$$S_J(x, P) = S_J(x) = \sum_{j=2}^{J} S^{(j}(x) = \sum_{j=2}^{J} P\left(G(x) \subset B(X_1, X_2, ..., X_j)\right).$$

The definition of depth provides a criterion to order the sample curves from the center-outward (from the deepest to the most extreme). This allows to extend order statistics to functional data. In particular, a median is a curve with maximum depth. An extensive study of the band depth and its properties can be seen in López-Pintado and Romo (2006).

The band depth finite-dimensional version is an alternative to the already existing definitions of depth that enjoys an interesting feature: it is computationally very fast and this makes it very convenient to deal with high-dimensional data. To visualize this particular case, each point x in $\mathbb{R}^d$ can be seen as a real function defined on the index set $\{1, 2, ..., d\}$, $x = (x(1), x(2), ..., x(d))$. Given points $x_1, x_2, ..., x_n$ in $\mathbb{R}^d$, let

$$R(x_1, x_2, ..., x_n) = \left\{ x \in \mathbb{R}^d : \min_{i=1,...,n} x_i(k) \leq x(k) \leq \max_{i=1,...,n} x_i(k) \right\} \tag{3}$$

be the $d-$dimensional interval with sides parallel to the axes and defined by the minimum and maximum coordinates of $x_1, x_2, ..., x_n$. The finite-dimensional band depth of any of these points x is

$$S_{n,J}(x) = \sum_{j=2}^{J} S_n^{(j}(x), \;\; J \geq 2, \tag{4}$$

where

$$S_n^{(j}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < ... < i_j \leq n} I\{x \in R(x_{i_1}, x_{i_2}, ..., x_{i_j})\}, \;\; j \geq 2, \tag{5}$$

is the proportion of sets (intervals) $R(x_{i_1}, x_{i_2}, ..., x_{i_j})$ defined by j different points $x_{i_1}, x_{i_2}, ..., x_{i_j}$ containing x.

López-Pintado and Romo (2006) propose also a more flexible definition of depth (the generalized band depth). The indicator function in the definition is replaced by the length of the set where the function is inside the corresponding band. For any function x in $x_1, ..., x_n$, let

$$\begin{aligned} A_j(x) &= A_{i_1,...,i_j}(x) = A(x; x_{i_1}, x_{i_2}, ..., x_{i_j}) \\ &= \left\{ t \in I : \min_{r=i_1,...,i_j} x_r(t) \leq x(t) \leq \max_{r=i_1,...,i_j} x_r(t) \right\},\ j \geq 2, \end{aligned}$$

be the set of points in the interval I where the function x is inside the band given by the observations $x_{i_1}, x_{i_2}, ..., x_{i_j}$. If λ is Lebesgue measure in $\mathbb{R}$, $\lambda_r = \frac{\lambda(A_j(x))}{\lambda(I)}$ is the proportion of time that x is inside the band. Now,

$$GS_n^{(j}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < ... < i_j \leq n} \lambda_r(A(x; x_{i_1}, x_{i_2}, ..., x_{i_j})),\ j \geq 2, \tag{6}$$

is a generalized version of $S_n^{(j}(x)$: if x is always inside the band, $\lambda_r(A_j(x))$ is one and this coincides with the previous definition of band depth.

For functions $x_1, ..., x_n$, the generalized band depth of one of these curves x is

$$GS_{n,J}(x) = \sum_{j=2}^{J} GS_n^{(j}(x),\ J \geq 2. \tag{7}$$

If $X_1, X_2, ..., X_n$ are independent copies of the process X generating the observations $x_1, ..., x_n$, the population version of these indexes is

$$GS^{(j}(x) = E\ \lambda_r(A(x; X_1, X_2, ..., X_j)),\ j \geq 2,$$

and

$$GS_J(x) = \sum_{j=2}^{J} GS^{(j}(x) = \sum_{j=2}^{J} E\lambda_r(A(x; X_1, X_2, ..., X_j)),\ J \geq 2.$$

The band depth is less adaptive than its generalized version and it is more depending on the curves shape. Another important difference between both definitions is their behavior for curves leaving the sample center only for a short interval, i.e., remaining in the interior of the sample most of the time but taking extreme values in short intervals: for these curves the generalized band depth can still be large whereas the band depth takes small values.

In the multivariate case, the generalized band depth of a point x in $\mathbb{R}^d$ is the proportion of its coordinates inside the bands (intervals) given by j different

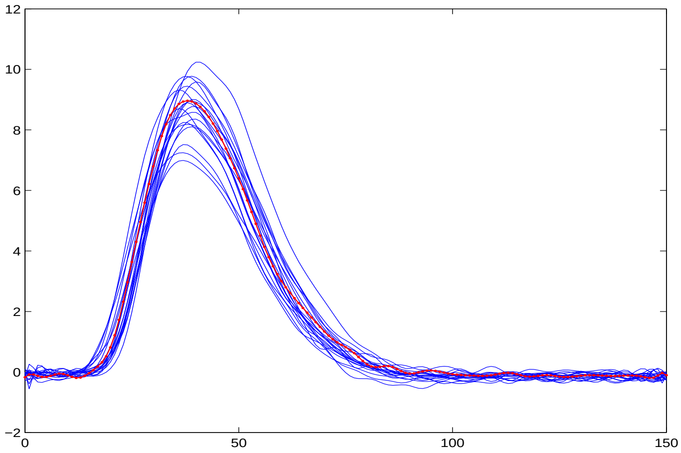

Fig. 2. *Deepest function (in red) for a sample of twenty curves.*

points in the sample:

$$GS_n^{(j}(x) = \binom{n}{j}^{-1} \sum_{i_1<\ldots<i_j} \frac{1}{d}\sum_{k=1}^{d} I\left\{\min_{r=i_1,\ldots,i_j} x_r(k) \le x(k) \le \max_{r=i_1,\ldots,i_j} x_r(k)\right\}$$

Thus, for example, with $j = 2$,

$$GS_n^{(2}(x) = \binom{n}{2}^{-1} \sum_{i_1<i_2} \frac{1}{d}\sum_{k=1}^{d} I\left\{x(k) \in seg(x_{i_1}(k), x_{i_2}(k)\right\}$$

where $seg(x_{i1}(k), x_{i2}(k))$ is the segment defined by the points $x_{i1}(k), x_{i2}(k)$.

Throughout the paper we will use the generalized band depth with $J = 2$ (López-Pintado and Romo (2006) provide evidence showing that the order induced in the sample is very stable with respect to growing J).

Figure 2 gives the deepest curve with the generalized band depth for twenty functions corresponding to the force exerted on a meter during a brief pinch by the thumb and forefinger (see Ramsay and Silverman, 2005). This curve describes adequately the center of the observations.

The remaining of the paper is organized as follows. In the next section, trimmed regions and contours are defined through the idea of depth. Moreover, it is proved that the trimmed regions constructed with the band depth verify the properties in Zuo and Serfling (2000) for finite-dimensional observations. The notion of central region is introduced in the third section, where some of their estimators are analyzed. Next section proposes a scale curve for functional data that allows to measure and represent the dispersion of a set of functions. Finally, all these nonparametric techniques are applied to several real data sets.

2 $\alpha-$trimmed functional regions

Liu et al. (1999) introduced several depth based inferential tools for multivariate data. Zuo and Serfling (2000) studied the structural properties of regions and contours for statistical depth functions. Next we extend them to functional observations. Let $X_1, ..., X_n$ be independent and identically distributed stochastic processes taking values in $C(I)$ with distribution P. Let $D(\cdot, P)$ be a functional depth and let $D(\cdot, P_n)$ be the corresponding sampling version. We will denote them by $D(\cdot)$ and $D_n(\cdot)$, respectively.

Definition 1 *Let $D(\cdot)$ be a functional depth and let $\alpha \geq 0$. The $\alpha-$trimmed region is*

$$D^\alpha = \{x \in C(I) : \; D(x) \geq \alpha\}$$

and the $\alpha -$ contour is

$$\partial D^\alpha = \{x \in C(I) : \; D(x) = \alpha\}.$$

The sample versions are

$$D_n^\alpha = \{x \in C(I) : \; D_n(x) \geq \alpha\}$$

and

$$\partial D_n^\alpha = \{x \in C(I) : \; D_n(x) = \alpha\},$$

respectively.

The next proposition provides the properties of the $\alpha-$trimmed region S^α for the finite-dimensional version of the band depth.

Theorem 2 *Let F be an absolutely continuous distribution in $\mathbb{R}^d$ with symmetric marginal distributions. Then:*

i. *$S^\alpha(F_{AX+b}) = A * S^\alpha(F_X) + b$ and $S_n^\alpha(F_{AX+b}) = A * S_n^\alpha(F_X) + b$, where A is a diagonal and invertible matrix and $b \in \mathbb{R}^d$.*

ii. *S^α is connected, i.e., it cannot be expressed as the union of two nonempty sets A and B, such that $\overline{A} \cap B = \varnothing$ and $A \cap \overline{B} = \varnothing$, where $\overline{A}$ and $\overline{B}$ are the closure of A and B, respectively.*

iii. *S^α and S_n^α are nested: for $\alpha_1 \geq \alpha_2$, $S^{\alpha_1} \subset S^{\alpha_2}$ and $S_n^{\alpha_1} \subset S_n^{\alpha_2}$.*

iv. *S^α is compact.*

Proof. The band depth properties can be seen in López-Pintado and Romo (2006). Part *i* follows from the invariance of $S(\cdot)$ and $S_{n,J}(\cdot)$. If F is absolutely continuous with symmetric marginals, $S(\cdot)$ is monotone and this implies the second property. Part *iii* is straightforward from the definition and, finally, S^α compactness holds because is bounded since depth tends to zero in infinity, and is closed under absolutely continuous distributions. ∎

He and Wang (1997) and Zuo and Serfling (2000) analyze the central regions and contours asymptotic properties for multivariate data in a very general context. Also, Mizera and Volauf (2002) study properties of contours constructed with the halfspace depth. Next result provides almost sure consistency of S_n^α; it relies on Theorem 4.1 in Zuo and Serfling (2000).

Theorem 3 *Let F be an absolutely continuous distribution on $\mathbb{R}^d$. Then, for all $\epsilon > 0$, $\delta < \epsilon, \alpha \geq 0$, and $\alpha_n \to \alpha$,*

i. $S^{\alpha+\epsilon} \subset S_n^{\alpha_n+\delta} \subset S_n^{\alpha_n} \subset S_n^{\alpha_n-\delta} \subset S^{\alpha-\epsilon}$ a.s., for large enough n (uniformly if $\alpha_n \to \alpha \in [0, \alpha_0]$ uniformly as $n \to \infty$, for $\alpha_0 < 1$).

ii. $S_n^{\alpha_n} \to S^\alpha$ a.s., $n \to \infty$, if $P\left\{x \in \mathbb{R}^d : \;\; S(x) = \alpha\right\} = 0$. The convergence is uniform in α if $\alpha_n \to \alpha \in [0, \alpha_0]$ uniformly as $n \to \infty$, for $\alpha_0 < 1$.

Proof. The proof is analogous to that of Theorem 4.1 in Zuo and Serfling (2000) since S and S_n verify the hypotheses in that theorem: (C1) $S(x) \to 0$, as $\|x\|_\infty \to \infty$ and (C2) $\sup\limits_{x \in R^d} |S_n(x) - S(x)| \overset{a.s.}{\to} 0$. ■

The properties for functional regions are contained in the following result. D is any of the functional depths defined in Section one.

Theorem 4 *Let P be a probability distribution on $C(I)$. Then:*

i. $D^\alpha(P_{ax+b}) = a * D^\alpha(P_x) + b$ and $D_n^\alpha(P_{ax+b}) = a * D_n^\alpha(P_x) + b$, where a and b are functions in $C(I)$ and a is different from zero for all $t \in I$.

ii. D^α and D_n^α are nested: for $\alpha_1 \geq \alpha_2$, $D^{\alpha_1} \subset D^{\alpha_2}$ and $D_n^{\alpha_1} \subset D_n^{\alpha_2}$.

As in the finite-dimensional case, the first property follows from invariance of the functional depth D. Property *ii* is straightforward from definitions of D^α and D_n^α.

3 p-central functional regions

Besides $\alpha-$trimmed regions, it is possible to extend also central regions and use them to analyze relevant distribution properties as dispersion. Assume again that D is any of the depths defined in Section one.

Definition 5 *The p-central region is*

$$C_p = \bigcap_\alpha \left\{D^\alpha : P\left(D^\alpha\right) \geq p\right\}, \, 0 \leq p \leq 1,$$

i.e., is the smallest set determined by trimmed regions containing at least probability p.

In the finite-dimensional case, the p-central region is compact and connected; its boundary is the p-contour and is denoted by ∂C_p. Moreover, if the density function is different from zero in $\mathbb{R}^d$ then ∂C_p is the contour of $\left\{x \in \mathbb{R}^d : D(x) = \alpha_p\right\}$, where $P\left\{x \in \mathbb{R}^d : D(x) \geq \alpha_p\right\} = p$. Thus, ∂C_p can be considered as a quantile surface. Following Liu et al. (1999), if the distribution is absolutely continuous and the density function different from zero,

$$C_p = D^{\alpha_p},$$

where $P\left(D^{\alpha_p}\right) = p$.

An important question is that the sample trimmed regions D_n^α are not directly observed and they have to be approximated. Next, we propose different ways of estimating both the trimmed regions D^α and the p-central regions C_p. Given a sample, either of points in $\mathbb{R}^d$ or continuous functions $x_1, ..., x_n$ in $C(I)$, they can be ordered from the deepest to the less deep object. This provides a center-outward ordering of the sample. Let $x_{(1)}, ..., x_{(n)}$ be the ordered statistic, where $x_{(1)}$ is the deepest element and $x_{(n)}$ is the most extreme. When depth ties occur, we consider, for simplicity, the following ordering procedure: if $x_{i_1}, ...x_{i_k}$ have the same depth, where $i_1 < i_2 < ... < i_k$, and there are exactly j sample points with larger depth, we assign $x_{(j+1)}, x_{(j+2)}, ..., x_{(j+k)}$ to $x_{i_1}, ...x_{i_k}$, respectively.

3.1 Finite-dimensional case

Let $x_{(1)}, ..., x_{(r_\alpha)}$ be the points in the ordered sample with depth larger or equal than α ($D_n(x_i) \geq \alpha$), i.e., $D_n\left(x_{(1)}\right) \geq D_n\left(x_{(2)}\right) \geq .... \geq D_n\left(x_{(r_\alpha)}\right)$, where $D_n\left(x_{(r_\alpha)}\right) \geq \alpha$ and $D_n\left(x_{(r_\alpha+1)}\right) < \alpha$.

Liu et al. (1999) proposed to estimate D^α with the convex envelope of the points having depth larger or equal than α,

$$\widehat{D_n^\alpha} = \text{convex envelope}\left\{x_{(1)}, ..., x_{(r_\alpha)}\right\}.$$

For this estimation, it holds that $\widehat{D_n^{\alpha_1}} \subset \widehat{D_n^{\alpha_2}}$, if $\alpha_1 \geq \alpha_2$. The p-central region C_p is estimated by the convex envelope of the proportion p of deepest sample points

$$C_{n,p} = \text{convex envelope}\left\{x_{(1)}, ..., x_{(|np|)}\right\},$$

where

$$|np| = \begin{cases} np, & \text{if } np \text{ is integer} \\ 1 + [np], & \text{in any other case.} \end{cases} \tag{8}$$

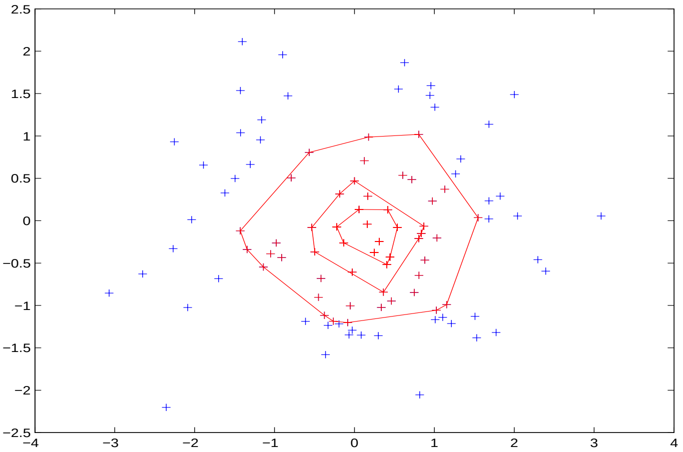

Fig. 3. *Estimated central regions $C_{n,p}$ for 100 points from a normal distribution with* $p = 0.1$, 0.2 and 0.5 (using *$S_{n,3}$ and GS*).

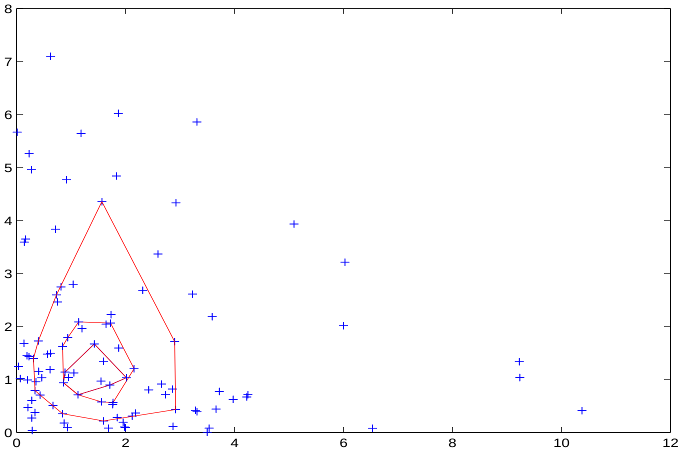

Fig. 4. *Estimated central regions $C_{n,p}$ for* 100 *points from a bivariate distribution with exponential marginals for* $p = 0.1$*,* 0.2 *and* 0.5 *with GS.*

$C_{n,p}$ approximates the sample p-central region and its boundary $\partial C_{n,p}$ is the p-contour or the p quantile surface. Figure 3 shows a sample of size 100 from a bivariate normal distribution with zero mean and covariance matrix $\Sigma = diag(2, 1)$. The 50% deepest points for $S_{n,3}$ and GS appear in red and the central regions are estimated by $C_{n,p}$ with $p = 0.1$, 0.2 and 0.5. In this example, the contours based on the band depth and the generalized band depth coincide. Figure 4 provides 100 points from a bivariate distribution with exponential marginals and the contours with the generalized band depth.

Trimmed and central regions could be also estimated through different strategies. Set estimates are an interesting alternative. A review of techniques developed in this area is Cuevas and Rodríguez-Casa (2003). For instance, we can consider balls centered at the sample points with depth larger than or equal to α as an estimation of the $\alpha-$trimmed region D^{α},

$$\overline{D}_n^{\alpha} = Ball(x_{(1)}, \varepsilon) \cup Ball(x_{(2)}, \varepsilon) \cup ... \cup Ball(x_{(r_\alpha)}, \varepsilon),$$

where $Ball(x, \varepsilon)$ is the closed ball in $\mathbb{R}^d$ centered in x with radius ε. A natural choice for the radius would be the minimum value for which $\overline{D}_n^{\alpha}$ is connected (Baíllo et al., 2000). By construction, $\overline{D}_n^{\alpha_1} \subset \overline{D}_n^{\alpha_2}$, if $\alpha_1 \geq \alpha_2$. The p-central region C_p could be estimated by $Ball(x_{(1)}, \varepsilon) \cup Ball(x_{(2)}, \varepsilon) \cup ... \cup Ball(x_{(\lfloor np \rfloor)}, \varepsilon)$.

3.2 Functional case

Next we propose approximations for the trimmed and central regions for functional observations. Let $x_1, ..., x_n$ be a sample of continuous functions in $C(I)$. Order them accordingly to the increasing depth D_n: $x_{(1)}, ..., x_{(n)}$. Let $x_{(1)}, ..., x_{(r_\alpha)}$ be the observations with depth larger or equal than α. We estimate the $\alpha-$trimmed region D^{α} as the band B delimited by the sample curves with depth larger than or equal to α,

$$\begin{aligned} \widetilde{D}_n^{\alpha} &= B(x_1, ..., x_{r_\alpha}) \\ &= \left\{ (t, y) \in I \times \mathbb{R} : \min_{i=1,\ldots,r_\alpha} \left\{ x_{(i)}(t) \right\} \leq y \leq \max_{r=1,\ldots,r_\alpha} \left\{ x_{(i)}(t) \right\} \right\}. \end{aligned}$$

Thus, the p-central region is

$$B_{n,p} = B\left(x_{(1)}, ..., x_{(\lfloor np \rfloor)} \right),$$

i.e., the band defined by the fraction p of deepest sample curves.

Figure 5 shows the central region $B_{n,0.15}$ delimited by the 15% of most central curves using the band depth $S_{n,3}$ and the generalized band depth. The data (see Ramsay and Silverman, 2005) are the hip angles in the sagittal plane when 39 children go through a gait cycle. Figure 6 provides the central region $B_{n,0.15}$ with $S_{n,3}$ for a different real example containing daily temperatures in 35 Canadian weather stations for one year (see Ramsay and Silverman, 2005). Both figures illustrate how the band amplitude adapts to the sample variability; this is an important property, because it reflects very adequately the structure of the set of functions.

4 A scale curve for functional data

In the multivariate context, the idea of depth has been used to describe graphically different properties of the underlying distribution, as, i.e., dispersion or

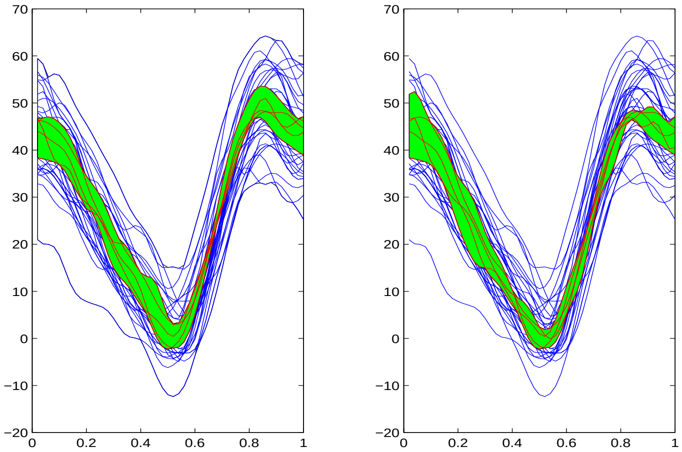

Fig. 5. *Hip angle in the sagittal plane when 39 children go through a gait cycle. Estimated central region (in green) with the 15% deepest curves for the band depth (left panel) and the generalized band depth (right panel).*

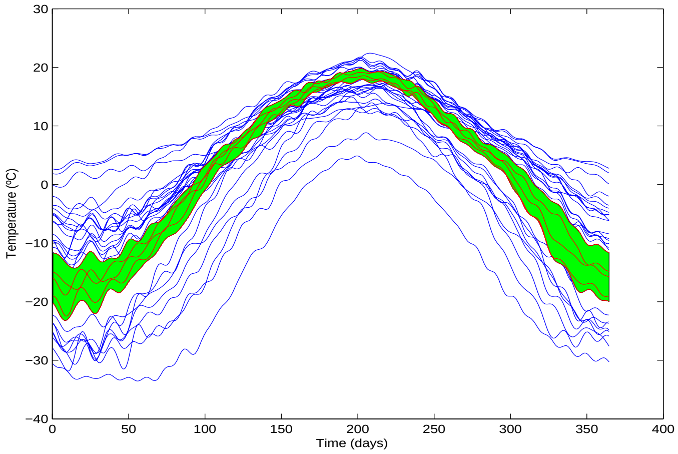

Fig. 6. *Daily temperature in 35 Canadian weather stations for one year. Estimated central region (in green) given by the 15% deepest curves for S_3.*

kurtosis. In Rousseeuw et al. (1999) and Liu et al. (1999), the box-plot is extended to multivariate observations using the central region. The scale curve (Liu et al., 1999) allows to measure and visualize the dispersion of a sample in $\mathbb{R}^d$. Any of these concepts can be now extended to functional observations. Next we generalize and apply the scale curve to functional data.

Definition 6 *The scale curve is the volume of the p-central region, $sc(p) = vol(C_p)$.*

In the finite-dimensional case, the curve $sc(p)$ can be estimated through the estimates of C_p; for example, the convex envelope $C_{n,p}$ of the deepest points. The scale curve for functional data that we propose is based on the central region $B_{n,p}$ and is the area of the band delimited by the proportion p of deepest functions.

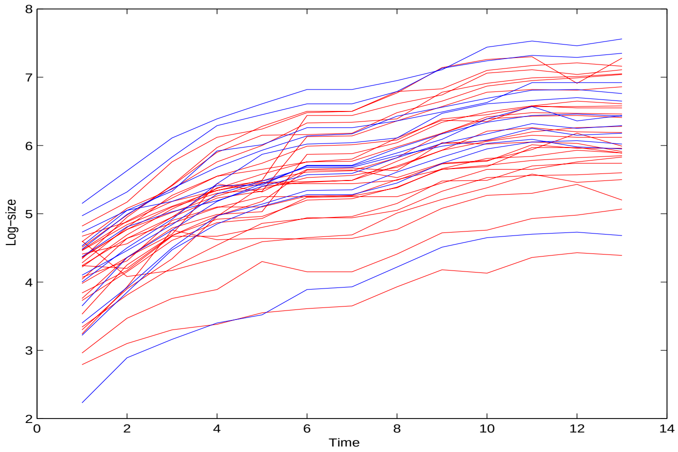

Fig. 7. *Size logarithm for two groups of trees. Blue curves correspond to a normal environment and red ones to an ozone enriched environment.*

Definition 7 *The scale curve of a set of functions $x_1, ..., x_n$ in $C(I)$ is*

$$A(p) = area\left\{(t,y) \in I \times \mathbb{R} : \min_{i=1,\ldots,|np|}\left\{x_{(i)}(t)\right\} \leq y \leq \max_{i=1,\ldots,|np|}\left\{x_{(i)}(t)\right\}\right\} =$$

$$= area\left\{B(x_{|1|}, ..., x_{|np|})\right\} = area\left\{B_{n,p}\right\}.$$

Thus, $A(p)$ is the area of the band delimited by the $|np|$ most central curves. The scale curve measures how the p-central region expands when p grows and is characterized by the speed of depth decreasing. From the computational point of view, it is convenient to rewrite $A(p)$ as

$$A(p) = \int_I \left(\max_{i=1,\ldots,|np|} x_{(i)}(t) - \min_{i=1,\ldots,|np|} x_{(i)}(t)\right) dt.$$

Next, we calculate and analyze the scale curve $A(p)$ for several real functional data sets. In all cases we have used $S_{n,3}$.

Figure 7 shows the growth of 39 trees. Twenty-nine of them live in an ozone enriched environment (represented in red) and the remaining ones are in a normal atmosphere (the blue ones). We consider the logarithm of the size (product of height and squared diameter of the trees) measured in thirteen different days (Diggle et al., 1994). The corresponding scale curves can be seen in Figure 8. The blue one corresponds to normal atmosphere and the red curve describes dispersion in an ozone enriched ambience.

For p smaller than 0.8, the scale curve $A(p)$ for the trees in a normal atmosphere is smaller than the curve for trees in an ozone enriched ambience. This means that dispersion is similar for central data; however, for p larger than 0.8, the dispersion of the ozone trees grows suddenly taking larger values. This reflects the presence of outliers. Looking back to Figure 7, we can find

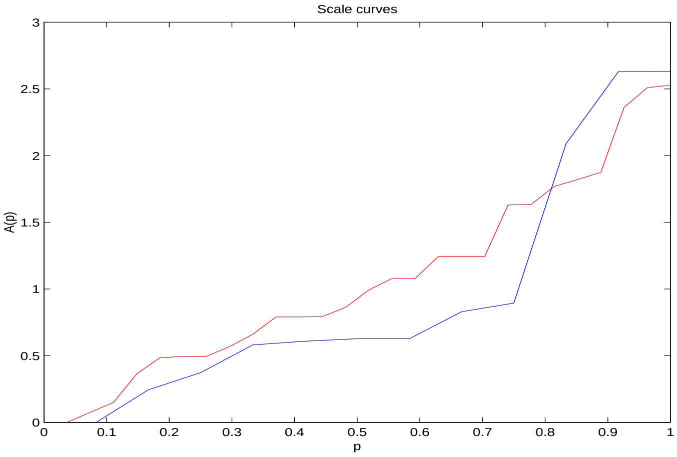

Fig. 8. *Scale curves for the trees growth data. The blue curve corresponds to normal environment and the red one to a ozone enriched atmosphere.*

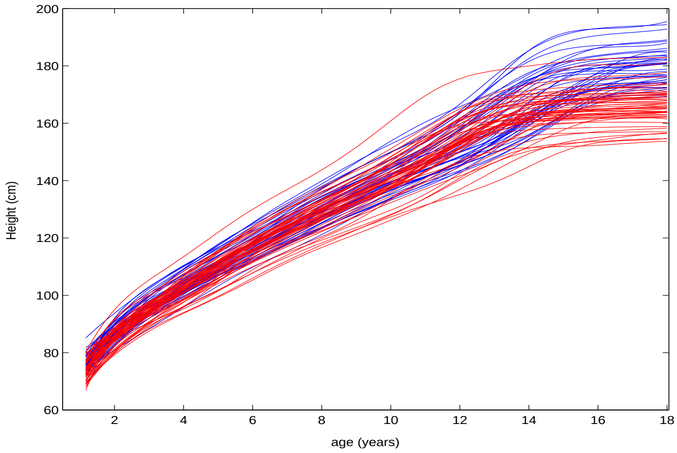

Fig. 9. *Heights of fifty four girls (in red) and thirty nine boys (in blue).*

an extreme blue curve.

The second real data set corresponds to the growth curves of thirty nine boys and fifty four girls measured at different instants from 1 to 18 years (Ramsay and Silverman, 2005). There are twenty nine values per child and they have been smoothed with a B-spline basis (Figure 9). The scale curves for these functions can be seen in Figure 10. Dispersion for girls is smaller than for boys when $p < 0.35$ and larger for $p > 0.35$. Moreover, the derivative of the scale curve reflects the dispersion rate of change with p; this rate of change is larger, except at the beginning, for girls than for boys.

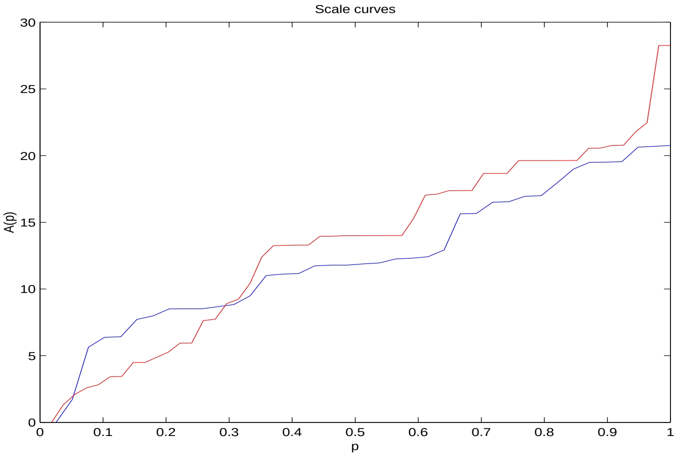

Fig. 10. *Scale curve for girls heights (in red) and for boys (in blue).*

5 Conclusions

We have proposed a robust nonparametric methodology for functional data based on the idea of depth. The depth ordering of the curves leads to extend to functions some finite-dimensional inference concepts and techniques. Thus, we have introduced the notions of trimmed and central regions for functions and established their properties. They allow to analyze the structure of a collection of curves. Also, we have defined a scale curve for functional data that provides a visual description of the sample data variability. All these tools have been applied to several real data sets. An important characteristic of all these techniques is that they are computationally very fast and can be also used for very high-dimensional observations.

References

Baíllo, A., Cuevas, A., Justel, A., 2000. Set estimation and nonparametric detection. The Canadian Journal of Statistics 28, 765–782.

Cuevas, A., Rodríguez-Casa, A., 2003. Recent Advances and Trends in Nonparametric Statistics. Eds. M.G. Akritas and D.N. Politis, Elsevier, pp. 251–264.

Fraiman, R., Muniz, G., 2001. Trimmed mean for functional data. Test 10, 419–440.

He, J., Wang, G., 1997. Convergence of depth contours for multivariate datasets. The Annals of Statistics 25, 495–504.

Liu, R., 1990. On a notion of data depth based on random simplices. The Annals of Statistics 18, 405–414.

Liu, R., Parelius, J., Singh, K., 1999. Multivariate analysis by data depth:

Descriptive statistics, graphics and inference. The Annals of Statistics 27, 783–858.

López-Pintado, S., Romo, J., 2006. On the concept of depth for functional data. working paper. Universidad Carlos III de Madrid.

Mahalanobis, P., 1936. On the generalized distance in statistics. Proceedings of the National Academy of Science of India 12, 49–55.

Mizera, I., Volauf, M., 2002. Continuity of halfspace depth contours and maximum depth estimators: diagnostics of depth related methods. Journal of Multivariate Analysis 83, 365–388.

Oja, H., 1983. Descriptive statistics for multivariate distributions. Statistics Probability Letter 1, 327–332.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis. Second Edition. Springer-Verlag. New York.

Rousseeuw, P., Ruts, I., Tukey, J., 1999. The bagplot: A bivariate boxplot. The American Statistician 53, 382–387.

Singh, K., 1991. A notion of majority depth. Unpublished document.

Tukey, J., 1975. Mathematics and picturing data. Proceedings of the 1975 International Congress of Mathematics 2, 523–531.

Vardi, Y., Zhang, C.-H., 2001. The multivariate L1-median and associated data depth. Proceedings of the National Academy of Science USA 97, 1423–1426.

Zuo, Y., 2003. Projection based depth functions and associated medians. The Annals of Statistics 31, 1460–1490.

Zuo, Y., Serfling, R., 2000. Structural properties and convergence results for contours of sample statistical depth functions. The Annals of Statistics 28, 483–499.