Computational Statistics and Data Analysis 55 (2011) 1679–1695

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

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).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.10.024

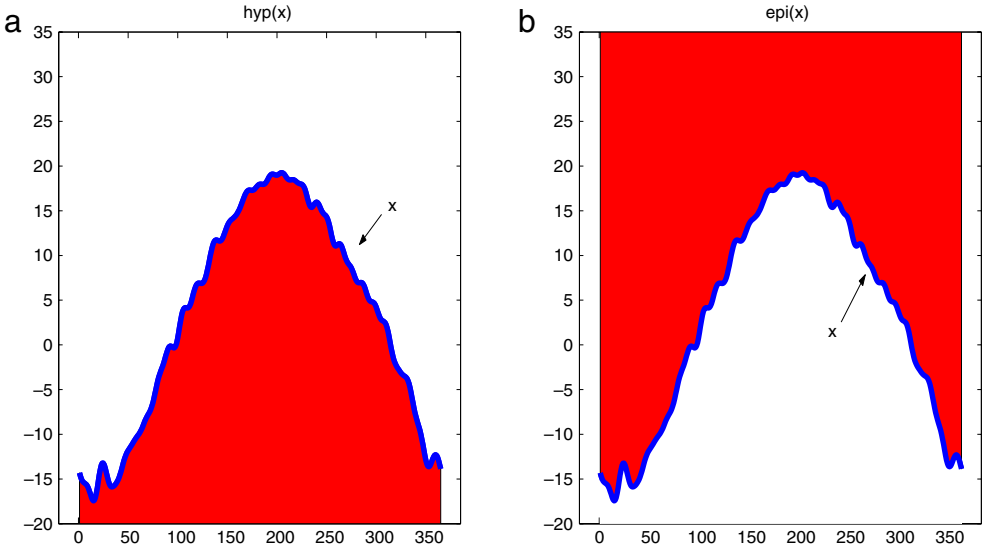

Fig. 1. (a) Hypograph and (b) epigraph of function x.

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

$$G_{1n}(x) = \frac{\sum_{i=1}^{n} I(G(x_i) \subset hyp(x))}{n} = \frac{\sum_{i=1}^{n} I(x_i(t) \leq x(t),\ t \in I)}{n},$$

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

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

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