Journal of Multivariate Analysis 142 (2015) 86–105

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Half-region depth for stochastic processes

James Kuelbs [a], Joel Zinn [b,*]

[a] Department of Mathematics, University of Wisconsin, Madison, WI 53706-1388, United States

[b] Department of Mathematics, Texas A&M University, College Station, TX 77843-3368, United States

ARTICLE INFO

Article history:
Received 3 February 2014
Available online 21 August 2015

AMS 2010 subject classifications:
primary 60F05
secondary 60F17
62E20

Keywords:
Depth
Consistency
Central limit theorems
Empirical processes

ABSTRACT

We study the concept of half-region depth, introduced in López-Pintado and Romo (2011). We show that for a wide variety of standard stochastic processes, such as Brownian motion and other symmetric stable processes with stationary independent increments tied down at 0, half-region depth assigns depth zero to all sample functions. To alleviate this difficulty we introduce a method of smoothing, which often not only eliminates the problem of zero depth, but allows us to extend the theoretical results on consistency in that paper up to the $\sqrt{n}$ level for many smoothed processes.

1. Introduction and some notation

A number of depth functions are available to provide an ordering of finite dimensional data, and more recently in [14] the interesting notion of half-region depth for stochastic processes was introduced. This depth applies to data given in terms of infinite sequences, as functions defined on some interval, and even in more general settings.

In this paper we focus on three items. The first is to show (see Section 2) that for many standard data sources this depth is identically zero, and hence one needs to be cautious when employing it. In particular, we will see sample continuous Brownian motion, tied down to be zero at $t = 0$ with probability one, assigns zero half-region depth to all functions $h \in C[0, 1]$, but we show this sort of behavior also holds for many other random processes widely used to model data in a variety of settings. A second item we examine is how the difficulty of zero half-region depth can be avoided, and fortunately in many situations smoothing the process by adding an independent real valued random variable Z with a density as in (28) (also see Proposition 4) changes things dramatically for half-region depth. In particular, it allows us to establish positivity for this depth and, as can be seen from Remark 4, the smoothed data remains a good approximation of the original input by taking $E(|Z|)$ small. Using Proposition 4 as in Remark 5, we also provide some sufficient conditions where smoothing is unnecessary for positive half-region depth.

The third item we consider involves limit theorems for the empirical half-region depth of these smoothed processes, and Theorem 1 is a basic consistency result with Theorem 2 and Corollary 6 providing some rates of convergence for this consistency. Moreover, a sub-Gaussian tail bound is obtained in Corollary 6. Theorem 3 implies a consistency result and $\sqrt{n}$-rates for half-region depth over all finite subsets of T of cardinality less than or equal to a fixed r. Hence, Theorem 3 extends the consistency result for random Tukey depth in [2,3] to half-region depth under weaker conditions, i.e. less

* Corresponding author.
E-mail addresses: kuelbs@math.wisc.edu (J. Kuelbs), jzinn@math.tamu.edu (J. Zinn).

http://dx.doi.org/10.1016/j.jmva.2015.07.012
0047-259X/

independence is used and the depth is computed using multi-dimensional marginals rather than those of one dimension. Now we turn to the notation used throughout the paper, and following that we indicate some additional details on our results and how they relate to other recent papers.

To fix some notation let $X := \{X(t) = X_t : t \in T\}$ be a stochastic process on the probability space $(\Omega, \mathcal{F}, P)$, all of whose sample paths are in $M(T)$, a linear space of real valued functions on T which we assume to contain the constant functions. To handle measurability issues, we also **always** assume that $h \in M(T)$ implies

$$\sup_{t\in T} h(t) = \sup_{t\in T_0} h(t) < \infty, \tag{1}$$

where T_0 is a fixed countable subset of T. Typical examples of $M(T)$ are the uniformly bounded continuous functions on T when T is a separable metric space, or the space of cadlag functions on T for T a compact interval of the real line. In either of these situations T_0 could be any countable dense subset of T. It should also be observed that since (1) holds on the linear space $M(T)$, then $h \in M(T)$ implies

$$\inf_{t\in T} h(t) = \inf_{t\in T_0} h(t) > -\infty \quad \text{and} \quad \|h\|_\infty \equiv \sup_{t\in T} |h(t)| = \sup_{t\in T_0} |h(t)| < \infty. \tag{2}$$

If $g, h\colon T \to \mathbb{R}$ and $S \subseteq T$, let $g \preceq_S h$ (resp., $g \succeq_S h$), denote that $g(t) \le h(t)$ (resp., $g(t) \ge h(t)$) for all $t \in S$. When $S = T$ we will simply write $g \preceq h$ (resp., $g \succeq h$). Then, for a function $h \in M(T)$, the half-region depth with respect to P is defined as

$$D(h, P) := D_{HR}(h, P) := \min(P(X \succeq h), P(X \preceq h)). \tag{3}$$

To simplify, we also will write $D(h)$ for $D(h, P)$ when the probability measure P is understood. Since $M(T)$ is a linear space with (1) and (2) holding, and the sample paths of the stochastic process X are in $M(T)$, we see for each $h \in M(T)$ that

$$\{X \preceq h\} = \{X \preceq_{T_0} h\} \quad \text{and} \quad \{X \succeq h\} = \{X \succeq_{T_0} h\}. \tag{4}$$

Thus the events in (3) are in $\mathcal{F}$ and the probabilities are defined.

Assume that $X, X_1, X_2, \ldots$ are i.i.d. copies of the process X defined on the probability space $(\Omega, \mathcal{F}, P)$ suitably enlarged, if necessary, such that all sample paths of each X_j are in $M(T)$. Then, the empirical half-region depth of $h \in M(T)$ based on the i.i.d. copies $X_1, \ldots, X_n$ is

$$D_n(h) = \min\left\{\frac{1}{n}\sum_{j=1}^n I(X_j \succeq h),\ \frac{1}{n}\sum_{j=1}^n I(X_j \preceq h)\right\}. \tag{5}$$

Throughout this paper to be certain the half-region depth is not degenerate at zero the smoothing we use is as in Proposition 4. However, the reader may care to notice that in Theorems 1 and 2 we actually assume more on the density $f_Z(\cdot)$, but those assumptions are only required to facilitate their proofs. The positivity of the half-region depth already holds under the weaker assumptions on $f_Z(\cdot)$ of Proposition 4. Other forms of smoothing may also be beneficial when seeking to avoid the problem of the depth being degenerate at zero, and some work is currently being done in this direction. To deal with the 0-depth problem López-Pintado and Romo [14] consider another depth, which they call modified half-region depth (see the definition below). There the depth itself is changed so as to be less restrictive and non-degenerate at zero, whereas here we retain the depth, but apply it to data which has been smoothed. One reason which motivates our choice, at least for us, is that there are examples where the ordering produced by modified half-region depth produces multiple medians, contrary to what one would intuitively expect. Furthermore, half-region depth typically orders the original paths or suitably smoothed paths in these examples so as to identify the intuitive median as being the unique median. To make this more precise, we consider the following simple examples.

In the first two examples $T = [0, 1]$, $\rho(\cdot)$ denotes Lebesgue measure on T, and we assume the sample functions of the stochastic process $\{Y(t) : t \in T\}$ are jointly measurable in (t, ω) with respect to Lebesgue measure on T and the probability $P = \mathcal{L}(Y)$. Then, if $h(\cdot)$ is a Lebesgue measurable function on T, the ρ-modified half-region depth of $h(\cdot)$ is

$$MD(h, P, \rho) = \min\left[\int_T P(h(t) \le Y(t))d\rho(t),\ \int_T P(h(t) \ge Y(t))d\rho(t)\right].$$

Example 1. If $\sup_{t\in[0,1]} |Y(t)| \le \lambda < \infty$ and $Y(t)$ has continuous distribution function for all $t \in [0, 1]$, then modified half-region depth based on Lebesgue measure on $[0, 1]$, never has a unique median. For any subset $A \subset [0, 1]$ with measure $1/2$, one considers $h_A := \lambda(2I_A - 1)$. Then, for each $t \in [0, 1]$, we have $P(Y(t) = -\lambda) = 0$ and that

$$P(Y(t) \le h_A(t)) = I_A(t) + I_{A^c}(t)P(Y(t) = -\lambda) = I_A(t).$$

Similarly, $P(Y(t) \ge h_A(t)) = I_{A^c}(t)$, and therefore the modified half-region depth of h_A is $1/2$. Since, $1/2$ is the maximal value of this depth when the distribution function of $Y(t)$ is continuous for all t, h_A is a median. In particular, among these medians we have $h_{(0,1/2)} \cdot h_{(1/2,1)} = 0$, and if the distribution of Y is symmetric enough around the zero function, neither of these functions seems an intuitive median. Furthermore, if we smooth the process Y as in Proposition 4, then the half-region depth of the smoothed process is positive, but unless we know more about the Y process it is still hard to determine the median for this half-region depth. Our next two examples are more specific, and allow us to make such determinations.