Journal of Multivariate Analysis 146 (2016) 325–340

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

An angle-based multivariate functional pseudo-depth for shape outlier detection

Sonja Kuhnt [a], André Rehage [b,*]

[a] Department of Computer Science, Dortmund University of Applied Sciences and Arts, 44227 Dortmund, Germany
[b] Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany

ARTICLE INFO

Article history:
Received 31 January 2015
Available online 3 November 2015

AMS 2010 subject classifications:
60G15
62H30
62P99
68-04

Keywords:
Bootstrap
Data depth
Functional data
Robust estimate
Shape outlier detection

ABSTRACT

A measure especially designed for detecting shape outliers in functional data is presented. It is based on the tangential angles of the intersections of the centred data and can be interpreted like a data depth. Due to its theoretical properties we call it functional tangential angle (FUNTA) pseudo-depth. Furthermore we introduce a robustification (rFUNTA). The existence of intersection angles is ensured through the centring. Assuming that shape outliers in functional data follow a different pattern, the distribution of intersection angles differs. Furthermore we formulate a population version of FUNTA in the context of Gaussian processes. We determine sample breakdown points of FUNTA and compare its performance with respect to outlier detection in simulation studies and a real data example.

1. Introduction

In every statistical analysis one should check the data for unusual, surprising observations. In one or two dimensions this might be done graphically. However, in higher dimensions or complex data structures, outlier detection methods are needed [4].

The analysis of outliers in functional data [32,14,20,6] is an interesting and recent field. Functional data occur when measurements at each observational unit are taken densely, mostly over time. The inherent smoothness of the measurements is used to approximate the data vector (or matrix) by a function. As in a regression setup (see e.g. [24]), in functional data it can be distinguished between different types of outliers. For example, magnitude and shape outliers are mentioned in [27]. A finer classification is revisited in [21]: Shift outliers are curves which are shifted away from the bulk of the data, whereas amplitude outliers are curves with a similar shape but differing scale. Furthermore the authors differentiate between persistent and isolated outliers. However, an isolated outlier with e.g. a larger scale than the remaining data might also be seen as a global (persistent) shape outlier. We focus on shape outliers, since shift and magnitude outliers can often be detected with well-known univariate approaches. A popular approach to apply a centre-outward ordering to multivariate as well as functional data is the concept of data depth. Several outlier identification approaches based on functional depth measures exist [9,12,27], but they are not specifically designed to detect shape outliers. See for example Fig. 1, where the dashed functions have the same modified band depth on each of the panels. In the left panel the solid

* Corresponding author.
E-mail addresses: sonja.kuhnt@fh-dortmund.de (S. Kuhnt), rehage@statistik.tu-dortmund.de (A. Rehage).

http://dx.doi.org/10.1016/j.jmva.2015.10.016
0047-259X/

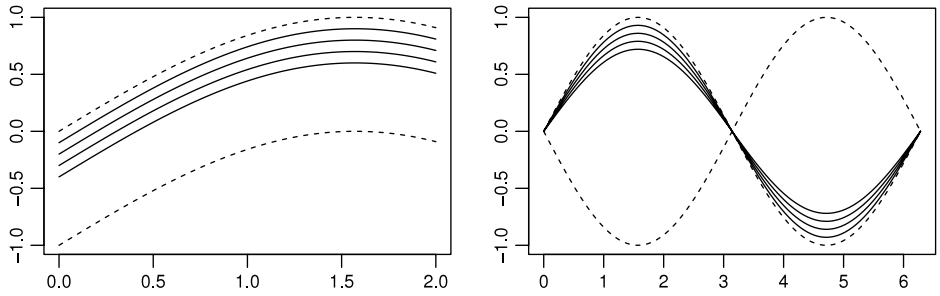

Fig. 1. Toy data examples where modified band depth [27] yields implausible results.

functions have a larger depth which is sensible for magnitude outlier detection, but not for shape outlier detection. In the right panel one of the dashed functions follows an obviously different pattern than the rest, but that is not reflected in the modified band depth. Recently, a graphical approach [3] has been proposed to overcome this issue. We try to remedy this shortcoming more directly by introducing a measure depending on the angles in the intersections of one observation with the others, which we call FUNTA pseudo-depth or to put it briefly: FUNTA. Similar to the location-slope graph depth [29], FUNTA depends on functions and their derivatives, but restricts the analysis to the intersections.

As in the non-functional case, one encounters not only univariate, but also multivariate data. Literature regarding the so-called multi-functional data analysis emerged recently, expanding methods like linear models [32], partial linear modelling [2], additive prediction and boosting [15] as well as linked applications like variable selection [8] to the multi-functional context. Besides, the concept of data depth has been extended to multi-functional situations: One way to deal with multivariate functional data is to compute one of the numerous multivariate data depths for each time point. This has been conducted for univariate and multivariate functional data in [21] using the skew-adjusted projection depth and a heat map. While in the univariate situation outliers could be found easily, the multivariate situation has turned out to be more ambiguous. Hence, the authors use a further graphical representation to clarify multivariate functional outliers in a dataset, called the centrality-stability plot. On the other hand, multivariate functional depths have been constructed to overcome this issue, for example the multivariate functional halfspace depth [7,22], the simplicial band depth for multivariate functional data [28] and a modification using weighting variance–covariance operators [5]. As a competitor to those depths, we introduce a multivariate version of FUNTA along with the scaled average pairwise interangular distance in this paper.

An alternative framework for outlier detection is described in [14]: In functional nonparametric unsupervised classification, the outlying and non-outlying data can be considered as two (or more) classes. The advantage of our approach is that the number of classes is irrelevant, hence the outliers do not need to be homogeneous in any way.

It is well-known that "for functional data the information contained in the *shape* of the curves matters a great deal" [20, p. 5]. Thus it is surprising that there are barely any depth measures focusing on the curves' shape. Certainly, depth measures were intended to determine a multivariate median and outlier detection is in various aspects different to median detection. Particularly in functional data situations we emphasise that a shape outlier is not necessarily the "least median" observation. To improve the ability of a measure to detect shape outliers, any information concerning the location should be removed from the data beforehand. In chemometrics, this procedure is known as the baseline correction (see e.g. [21]). Therefore we will mainly refer to "centred functions" in the following section.

The paper is organised as follows. In Section 2 we introduce two different notions of the functional tangential angle pseudo-depth, both univariate and multivariate. A population version is stated in the case of Gaussian processes. We also discuss theoretical properties of FUNTA, particularly breakdown points. We compare the ability to detect outliers in different artificial data situations in Section 3. Besides the two versions of FUNTA we choose h-modal depth, modified band depth, the integrated depth and the multivariate functional halfspace depth as competitors. A real data example from linguistics is analysed in Section 4. Conclusions and projections for further research are given in Section 5.

2. Methodology

In this section we present the FUNTA measurement based on functional tangential angles and then introduce robustified, multivariate and population versions. Computational issues are discussed and sample breakdown properties derived.

2.1. *Sample versions of* FUNTA *pseudo-depth*

Before we define FUNTA in a slightly different notation than in [35], we state that by a sample of "centred functions" we mean that $\int x_i(t)dt = \int x_j(t)dt \; \forall\, i, j$.

Definition 1 (*Functional Tangential Angle Pseudo-Depth:* FUNTA)**.** Let $\tilde{x}(t), x_1(t), \ldots, x_n(t), t \in \mathcal{T} = [a, b]$, be a sample of real centred differentiable functions, where $\tilde{x}(t)$ has a finite number of intersections with $x_i(t), i = 1, \ldots, n$, denoted by