Statistical Methodology 20 (2014) 27–39

Contents lists available at ScienceDirect

Statistical Methodology

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

The deepest point for distributions in infinite dimensional spaces

Anirvan Chakraborty *, Probal Chaudhuri

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute 203, B. T. Road, Kolkata - 700108, India

ARTICLE INFO

Article history:
Received 14 November 2012
Received in revised form
22 April 2013
Accepted 28 April 2013

Keywords:
Breakdown point
Coordinatewise median
Functional depths
Spatial depth
Spatial median
Strong consistency

ABSTRACT

Identification of the center of a data cloud is one of the basic problems in statistics. One popular choice for such a center is the median, and several versions of median in finite dimensional spaces have been studied in the literature. In particular, medians based on different notions of data depth have been extensively studied by many researchers, who defined median as the point, where the depth function attains its maximum value. In other words, the median is the deepest point in the sample space according to that definition. In this paper, we investigate the deepest point for probability distributions in infinite dimensional spaces. We show that for some well-known depth functions like the band depth and the half-region depth in function spaces, there may not be any meaningful deepest point for many well-known and commonly used probability models. On the other hand, certain modified versions of those depth functions as well as the spatial depth function, which can be defined in any Hilbert space, lead to some useful notions of the deepest point with nice geometric and statistical properties. The empirical versions of those deepest points can be conveniently computed for functional data, and we demonstrate this using some simulated and real datasets.

1. Introduction

For a univariate probability distribution, the median is a well-known and popular choice of its center. It has several desirable statistical properties, which include equivariance under monotone transformations, asymptotic consistency under very general conditions, and high breakdown point.

* Corresponding author. Tel.: +91 9477377387.
E-mail addresses: vanchak@gmail.com (A. Chakraborty), probal@isical.ac.in (P. Chaudhuri).

1572-3127/$ – see front matter
http://dx.doi.org/10.1016/j.stamet.2013.04.004

The concept of the median has been extended in several ways for probability distributions in finite dimensional Euclidean spaces (see, e.g., [8,22] for some reviews). The median can also be defined as the point in the sample space with the highest depth value with respect to an appropriate depth function (see, e.g., [9,15,22,26]). Due to the recent advances in technology and measurement devices, statisticians frequently have to analyze data for which the number of variables is much larger than the sample size. Such data can be conveniently viewed as random observations from probability distributions in infinite dimensional spaces, e.g., the space of real-valued functions defined on an interval.

It turns out that many medians for finite dimensional probability measures do not extend in any natural and meaningful way into infinite dimensional spaces. On the other hand, an extension of the well-known spatial median (see, e.g., [4]) into Banach spaces was studied in [12,24]. There has been some recent work on developing depth functions for probability measures in function spaces. The integrated data depth [10], the band depth [16] and the half-region depth [17] have been defined for data in the space of real-valued continuous functions on an interval on the real line. The authors of those papers have used these depth functions to construct depth based trimmed means, to find central and extreme curves for several real datasets, and also to construct test procedures based on depth based ranking. Recently, the authors of [23] have used the deepest point based on the band depth in the construction of boxplots for functional data. In Section 2, we shall critically investigate the deepest points associated with some of the functional depths. In particular, we show that both of the band depth and the half-region depth assign zero value to the center of symmetry for a large class of symmetric stochastic processes on $C[0, 1]$. Thus, there is no meaningful deepest point associated with either of these two depth functions for such stochastic processes. However, the integrated data depth, the spatial depth, and some modified versions of the band depth and the half-region depth yield statistically meaningful notions of deepest points. In Section 3, we shall demonstrate the empirical deepest points associated with these depth functions using some simulated and real functional data. It is shown that the empirical deepest points associated with the integrated data depth, the spatial depth and the modified versions of the band depth and the half-region depth can be easily computed for any given functional data.

In the univariate setting, one of the main motivations for considering the median is its robustness against outlying observations. According to the traditional measures of robustness like breakdown point, the median is a more robust estimator of location than the mean. In an interesting paper [21], the author considered an alternative measure of robustness assuming that all the observations remain bounded. Using that measure, the author of [21] obtained some counter-intuitive results regarding the robustness of the median and the mean (see also [18]). In particular, it was shown that the mean may be more robust than the median under certain conditions. Unlike the univariate sample median, many multivariate depth based medians fail to achieve 50% breakdown point (see, e.g., [22] for a review). In fact, in $\mathbb{R}^d$ for $d \geq 2$, the half-space median has a breakdown point of 1/3, while that for the simplicial median is atmost $1/(d+2)$ (see, e.g., [8] for a brief review and relevant references). In Section 4.1, we will consider the breakdown points of the empirical deepest points based on the integrated data depth, the spatial depth and the modified versions of the band depth and the half-region depth. We show that the empirical deepest point associated with the spatial depth has 50% breakdown point. We also show that there is an empirical deepest point associated with each of the integrated data depth, the modified band depth and the modified half-region depth that has 50% breakdown point.

The strong consistency of the empirical versions of most of the medians for finite dimensional data is well-known in the literature. However, for infinite dimensional data, the situation is much more complex. In some cases, like the medians based on the band depth and the half-region depth, strong consistency has been proved under the assumption that the empirical medians along with the unique population median lie in a fixed equicontinuous set (see [16,17]). For the empirical spatial median in infinite dimensions, the convergence has been shown to hold only in a weak sense, namely, for continuous real linear functions of the estimator (see, e.g., [5,11]). Although the author of [5] proved strong consistency of the empirical spatial median in the norm topology of the Hilbert space $L_2(\mathbb{R})$, the result requires extremely strong assumptions on the underlying stochastic process, which include boundedness and differentiability of sample paths. Thus, this result is not valid for many important processes including the standard Brownian motion and fractional Brownian motions. Although the