Computational Statistics and Data Analysis 98 (2016) 19–30

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Exact computation of the halfspace depth

Rainer Dyckerhoff *, Pavlo Mozharovskyi

Institute for Econometrics and Statistics, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany

ARTICLE INFO

Article history:
Received 24 November 2014
Received in revised form 12 December 2015
Accepted 17 December 2015
Available online 29 December 2015

Keywords:
Tukey depth
Exact algorithm
Projection
Combinatorial algorithm
Orthogonal complement

ABSTRACT

For computing the exact value of the halfspace depth of a point w.r.t. a data cloud of n points in arbitrary dimension, a theoretical framework is suggested. Based on this framework a whole class of algorithms can be derived. In all of these algorithms the depth is calculated as the minimum over a finite number of depth values w.r.t. proper projections of the data cloud. Three variants of this class are studied in more detail. All of these algorithms are capable of dealing with data that are not in general position and even with data that contain ties. As is shown by simulations, all proposed algorithms prove to be very efficient.

1. Introduction

In 1975 John W. Tukey suggested a novel way of ordering multivariate data. He proposed to order the data according to their centrality in the data cloud. To achieve this, he defined what is nowadays known as *halfspace depth* (also called Tukey depth or location depth). Motivated by his proposal many notions of data depth have been proposed in the last decades, e.g., the simplicial depth (Liu, 1988, 1990), the projection depth (Liu, 1992; Zuo and Serfling, 2000; Zuo, 2003; based on a notion of outlyingness proposed by Stahel, 1981; Donoho, 1982), and the zonoid depth (Koshevoy and Mosler, 1997).

Data depths have been applied in many fields of statistics, among others in multivariate data analysis (Liu et al., 1999), statistical quality control (Liu and Singh, 1993), classification (Mosler and Hoberg, 2006; Lange et al., 2014), tests for multivariate location and scale (Liu, 1992; Dyckerhoff, 2002), multivariate risk measurement (Cascos and Molchanov, 2007), and robust linear programming (Mosler and Bazovkin, 2014).

Consider a data cloud $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ with data points $\boldsymbol{x}_i \in \mathbb{R}^d$. Conditioned on it, a statistical depth function assigns to an arbitrary point $\boldsymbol{z} \in \mathbb{R}^d$ its degree of centrality, $\boldsymbol{z} \mapsto D(\boldsymbol{z}|\boldsymbol{X}) \in [0, 1]$.

The halfspace depth is determined as the smallest fraction of data points contained in a closed halfspace containing $\boldsymbol{z}$. This classical depth function, based on the ideas of Tukey (1975) and later developed by Donoho and Gasko (1992), is one of the most important depth notions and is historically the first one. It possesses a number of attractive properties, such as affine invariance, monotonicity on rays from any deepest point, quasiconcavity and upper semicontinuity, studied in Zuo and Serfling (2000), Dyckerhoff (2004) and Mosler (2013). The halfspace depth determines uniquely the empirical distribution (Koshevoy, 2002) and takes a finite number of values in the interval from 0 (for the points that lie beyond the convex hull of the data) to its maximum value (which depends on the data, but is 1/2 if all data points are different from $\boldsymbol{z}$), increasing by a multiple of $1/n$. By its nature the halfspace induced maximizers (the Tukey median) have a relatively high breakdown point.

* Correspondence to: Institut für Ökonometrie und Statistik, Universität zu Köln, 50923 Köln, Germany. Tel.: +49 2214704268; fax: +49 2214705084.
E-mail addresses: rainer.dyckerhoff@statistik.uni-koeln.de (R. Dyckerhoff), mozharovskyi@statistik.uni-koeln.de (P. Mozharovskyi).

http://dx.doi.org/10.1016/j.csda.2015.12.011
0167-9473/

The task of calculating the halfspace depth has a direct connection to the regression depth of Rousseeuw and Hubert (1999) and to the risk-minimizing separating hyperplane in classification. Further, computing the halfspace depth essentially coincides with the densest hemisphere problem (Johnson and Preparata, 1978), which is of non-polynomial complexity in (n, d). For this reason a great part of the literature on the halfspace depth considers its computational aspects. Below, we give an overview of the history of its exact (in Section 1.1) and approximate (in Section 1.2) computation.

1.1. Previous approaches to calculation of the halfspace depth

The idea of the halfspace depth has been introduced in a conference paper by Tukey (1975). A similar mechanism of cutting by hyperplanes (lines in the two-dimensional case) has also been used for a bivariate sign test by Hodges (1955). During the last decades, a variety of attempts have been made to compute the halfspace depth and its trimmed regions.

Rousseeuw and Ruts (1996) pioneered in exactly calculating the halfspace depth for bivariate data clouds and constructing its contours (Ruts and Rousseeuw, 1996a,b) by exploiting the idea of a circular sequence (Edelsbrunner, 1987). Here, the depth of a point is computed with complexity $O(n \log n)$, and a single depth region is constructed with complexity $O(n^2 \log n)$, both essentially determined by the complexity of the QUICKSORT procedure. Johnson et al. (1998) suggest to account for a small subset of points only when constructing the first l depth contours, which yields a better complexity for small l (algorithm FDC). The halfspace depth describes a data cloud by a finite number of depth contours. Miller et al. (2003) compute all these with complexity $O(n^2)$, by which the depth of a single point can be afterwards calculated with complexity $O(\log^2 n)$.

Rousseeuw and Struyf (1998) introduce an algorithm to compute the halfspace depth for $d = 3$ with complexity $O(n^2 \log n)$. They project points onto planes orthogonal to the lines connecting each of the points from $\boldsymbol{X}$ with $\boldsymbol{z}$ and then calculate the halfspace depth in these planes using the algorithm of Rousseeuw and Ruts (1996). Bremner et al. (2006) calculate the halfspace depth with a primal–dual algorithm by successively updating upper and lower bounds by means of a heuristic till they coincide. Bremner et al. (2008) design an output-sensitive depth-calculating algorithm that represents the task as two maximum subsystem problems for $d > 2$. The latter ones are then run in parallel.

An interesting issue of updating the depth when points are added continuously to the data set is handled by Burr et al. (2011) for bivariate depth and depth contours.

Kong and Mizera (2012) employ direction quantiles defined as halfspaces corresponding to quantiles on univariate projections and prove their envelope to coincide with the corresponding halfspace depth trimmed region. In exact computation of depth trimmed regions for $d > 2$, Mosler et al. (2009) pioneered for the zonoid depth (see also Mosler, 2002). Their idea to segment $\mathbb{R}^d$ into direction cones has been exploited in later algorithms for computing depth and depth regions, also for the halfspace depth. Hallin et al. (2010) establish a direct connection between multivariate quantile regions and halfspace depth trimmed regions. The multivariate directional quantile for a given direction corresponds to a hyperplane that may carry a facet of a depth trimmed region. More than that, the authors define a polyhedral cone containing all directions yielding the same hyperplane; the union of the finite set of all these cones fills the $\mathbb{R}^d$. So, by the breadth-first search algorithm a family of hyperplanes, each defining a halfspace, is generated, and the intersection of these halfspaces forms the halfspace depth trimmed region (see also Paindaveine and Šiman, 2012a,b). Paindaveine and Šiman (2012a) state that the average complexity of their algorithm is not worse than $O(n^d)$ and the worst-case complexity is of order $O(n^{d+1})$.

Based on an idea similar to Hallin et al. (2010), Liu and Zuo (2014) compute the halfspace depth exactly by using a breadth-first search algorithm to cover $\mathbb{R}^d$ and QHULL to define the direction cones. In a most recent article Liu (2014) suggests two more algorithms for the exact computation of the halfspace depth, which seem to be very fast. One of these algorithms, the so-called refined combinatorial algorithm, can be seen as a special case of the framework developed in the current paper.

1.2. Approximation of the halfspace depth

Even with the fastest available algorithms the exact calculation of the halfspace depth is very elaborate and amounts to exponential complexity in n and d. Therefore, one tries to save computation expenses by approximating the depth. Following Dyckerhoff (2004), the halfspace depth satisfies the weak projection property, i.e., it is the smallest achievable depth on all one-dimensional projections, and thus can be bounded from above by univariate depths.

Rousseeuw and Struyf (1998), when suggesting the algorithm computing the halfspace depth for $d = 3$, explored four algorithms differing in how the directions to project the data are generated. They propose to take a random subset of: (1) all lines connecting $\boldsymbol{z}$ and a point from $\boldsymbol{X}$, (2) all lines connecting two points from $\boldsymbol{X}$, (3) all lines normal to hyperplanes based on $\boldsymbol{z}$ and $d - 1$ pairwise distinct points from $\boldsymbol{X}$, (4) all lines normal to hyperplanes based on d pairwise distinct points from X, and claim the last variant to work best. Cuesta-Albertos and Nieto-Reyes (2008) suggest to generate directions uniformly on $\mathbb{S}^{d-1}$. This method proves to be useful in classification. Afshani and Chan (2009) present a randomized data structure keeping the approximated depth value in some range of deviations from its real value.

The latter work of Chen et al. (2013) determines the number of tries needed to achieve a required precision exploiting the third approximation method of Rousseeuw and Struyf (1998). The authors also present its generalization by projecting $\boldsymbol{z}$ and $\boldsymbol{X}$ onto affine spaces of dimension greater than one. They report the approximated depth values to be exact in most of the experiments and the approximation errors never to be larger than $2/n$.