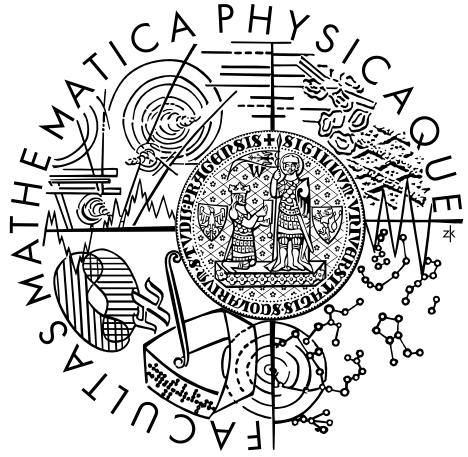
Charles University in Prague
Faculty of Mathematics and Physics

DOCTORAL THESIS



Ondřej Vencálek

Weighted Data Depth and Depth Based Discrimination

Department of Probability and Mathematical Statistics

Supervised by:	doc. RNDr. Daniel Hlubinka Ph.D.
Study programme:	Mathematics
Specialization:	Probability and Mathematical Statistics

Prague 2011

Rád bych poděkoval Danielu Hlubinkovi nejen za odborné rady, ale hlavně za vstřícný přístup po celou dobu mých studií. Vedení v jeho podání bylo směsí inspirujících otázek, drobných připomínek a velké míry povzbuzení při řešení nejen matematických problémů. Nakonec, byl to právě Daniel, kdo zprostředkoval první kontakt s mým současným působištěm v Olomouci.

Velký dík patří také mým rodičům. Jejich péče a láska ze mě udělaly to, čím dnes jsem; jejich podpora mi umožnila vystudovat.

V neposlední řadě chci poděkovat mé milované ženě Lucce, která se mnou byla prakticky denně po celé čtyři roky mých doktorských studií a sdílela se mnou trápení i radosti, které mi přinášela práce v oblasti matematické statistiky.

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague, August 15, 2011

Název práce: Vážená hloubka dat a diskriminace založená na hloubce dat

Autor: Ondřej Vencálek

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí disertační práce: doc. RNDr. Daniel Hlubinka Ph.D., Katedra pravděpodobnosti a matematické statistiky, Matematicko-fyzikální fakulta, Univerzita Karlova v Praze

Abstrakt: Hloubka dat je jedním z neparametrických nástrojů pro analýzu mnohorozměrných dat. Práce nově zavádí zobecnění poloprostorové hloubky, tzv. váženou hloubku dat. Vážená hloubka není obecně afinně invariantní, má však některé dobré vlastnosti, například že její centrální oblasti (oblasti s největší hloubkou) mohou být nekonvexní. Práce se dále zabývá možností aplikace metodologie hloubky dat v diskriminační analýze. Přehled klasifikátorů založených na hloubce dat je doplněn o návrh nového klasifikátoru, který je modifikací metody k nejbližších sousedů. Kvalita klasifikátorů je vyšetřována jak teoreticky (asymptotické vlastnosti), tak i v krátké simulační studii. V závěru je poukázáno na výhody, které lze získat použitím nově navržené vážené hloubky dat.

Klíčová slova: hloubka dat, neparametrické metody, mnohorozměrný, diskriminační analýza

Title: Weighted Data Depth and Depth Based Discrimination

Author: Ondřej Vencálek

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Daniel Hlubinka Ph.D., Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University Prague

Abstract: The concept of data depth provides a powerful nonparametric tool for multivariate data analysis. We propose a generalization of the well-known half-space depth called weighted data depth. The weighted data depth is not affine invariant in general, but it has some useful properties as possible nonconvex central areas. We further discuss application of data depth methodology to solve discrimination problem. Several classifiers based on data depth are reviewed and one new classifier is proposed. The new classifier is a modification of k-nearest-neighbour classifier. Classifiers are compared in a short simulation study. Advantage gained from use of the weighted data depth for discrimination purposes is shown.

Keywords: data depth, nonparametric methods, multivariate, discrimination

Contents

Why am I writing on this topic? Partly because picturing of data is important. Partly because, if present trends continue, an increasing fraction of all mathematicians will touch — or come close to touching — data during the next few decades.

John Tukey in his paper "Mathematics and the Picturing of Data", which started the development of data depth methodology (1975).

Introduction

The concept of data depth provides an important nonparametric approach to multivariate data analysis. The concept has been developed in the last twenty years. An impulse to the development of nonparametric methods was untenability of classical assumptions of parametric approach (similarly as in the univariate case). A very strong assumption of multivariate normal distribution is not satisfied in many cases. Nonparametric inference in the univariate case is based on data ordering. This ordering is very natural, because of natural (linear) ordering of real numbers. It enables us to define rank statistics. Also the notion of quantile (as one of the characteristics of univariate distribution) is based on this ordering.

Any extension to the higher dimension must deal with the problem of no natural ordering of real vectors. This makes problems when we are looking for generalization of quantile or rank statistics for multivariate distributions. Data depth concept provides one possible way how to order the multivariate data. We call this ordering as *central-outward ordering*. It should be noted that data depth provides quasi-ordering rather than ordering, because depth of some points in multidimensional space can be equal.

Basically, any function which provides "reasonable" central-outward ordering of points in multidimensional space can be considered as a depth function. This vague understanding of the notion of depth function led to the variety of depth functions, which have been introduced ad hoc since mid seventieth of the 20th century. The most important depth functions - halfspace depth, simplicial depth, zonoid depth and L_1 depth - are shortly introduced in Section 1.2. Other depth functions are mentioned as well.

The first depth function was introduced by Tukey in 1975. It is known as the halfspace depth and is still one of the most popular and the most widely used notions of data depth. Tukey's work started development of data depth concept. Historical overview should not missed many outstanding works, but the real milestone was the work by Zuo and Serfling "General notions of statistical depth function" published in 2000 ([56]). This work unified the theory of data depth by formulating a general definition of depth function. This definition states four main desirable properties of depth function. Section 1.1 is devoted to this general definition.

Data depth is very useful tool for nonparametric multivariate data inference. Basic tools such as descriptive characteristics of location, scale, skewness and kurtosis of multivariate distribution can be based on data depth. These characteristics were proposed in an outstanding article by Liu, Parelius and Singh in 1999 ([38]). The

paper deals also with visualizing these features via one-dimensional curves. More sophisticated methods based on data depth are reviewed in the article [50] by Serfling. Depth-based statistical procedures include tests of multivariate symmetry, diagnosis of non-normality, comparison of several distributions (for example tests of equal scales), outlier identification, statistical process control procedures, multivariate density estimation and some others. Section 1.3 presents several particularly interesting applications of data depth concept.

Chapter 1, which provides a detailed review of the data depth methodology, ends by some results and comments included in Section 1.4, that should clarify advantages and disadvantages of the data depth concept.

The research presented in Chapter 2 was motivated by some weak points of the halfspace depth. Central areas (areas including points with the highest depth) of the halfspace depth are always convex. We found this property undesirable, when we consider a distribution, whose density has nonconvex levelsets. Use of weights in the halfspace leads to a generalization of the halfspace depth, so called weighted halfspace depth. This newly proposed depth function allows the central regions not to be convex.

The concept of weighted data depth provides a broad class of possible weight functions, as can be seen in Section 2.2. Their basic properties are studied in Section 2.3. Choice of the weight function which determines weighted data depth is discussed in Sections 2.4 and 2.5. Computational aspects of the weighted data depth are discussed in Section 2.6. The last section of the second chapter illustrates differences between the halfspace depth and the weighted halfspace depth.

In the last ten years, much effort has been devoted to the application of data depth methodology in the area of discrimination. Chapter 3 deals with these applications. Methods, that have been proposed since 2000, are studied and their weak points are detected. Many classifiers based on data depth (mainly those introduced in Section 3.3) have problems when considered distributions differ in dispersion. This phenomenon can be explained by discrepancy between the (affine invariant) depth function and (non affine invariant) density function. Several solutions of this problem was proposed. Classifiers which do not suffer from the problem of distributions with different dispersions are introduced in Section 3.4, including a newly proposed classifier based on a modified k-nearest-neighbour method. A distinct way of discrimination is described in Section 3.5. Depth based classifiers are compared to the classical classifiers in a simulation study presented in Section 3.7. An advantage from the use of the weighted data depth introduced in Chapter 2 is shown there.

Chapter 1

Data depth

1.1 General definition of depth function and some basic concepts

Existence of broad variety of different "data depth" notions, that have been proposed ad hoc since mid seventieth of the 20th century, called for unified theory of data depth. Different notions of data depth have different more or less desirable properties. Probably the first work mapping systematically properties of some data depth notion was that by Liu ([36]). On the basis of these desirable properties Zuo and Serfling formulated the general definition of statistical depth function ([56]). Besides rather technical assumptions of nonnegativity and boundedness of a depth function, there are four main properties of a depth function. We list them here with a short discussion:

- *Affine invariance.* Informally said, the depth of a certain point $\boldsymbol{x}$ should not change when doing some affine transformation like shifting or rotating all points or changing the scale of measurements. Shortly said it should not depend on the underlying coordinate system. Affine invariance of depth function preserves affine equivariance of location estimators based on this depth.

- *Vanishing at infinity.* The depth of a point $\boldsymbol{x}$ should be very small when the point is far away from the others (when its norm approaches infinity). This is very natural requirement, which is similar to the requirement on probability density.

- *Maximality at centre.* If the probability distribution P is symmetric somehow, with the unique point that can be called the centre of symmetry, then this point should be the deepest point, that is the point, where the depth function attains its maximum value.

  Zuo and Serfling use the term "centre" to denote a point of symmetry. Several notions of symmetry can be considered. The authors considered three particularly important notions of multivariate symmetry: central, angular and halfspace symmetry. We add the notion of elliptical symmetry, which is considered in many applications. Broader discussion can be found in [13]. Serfling wrote an engaging overview of multivariate symmetry and asym-

metry [51], more theoretical results are presented in [57]. Distribution of a d-dimensional random vector $\boldsymbol{X}$ is said to be:

- *elliptically symmetric* about $\boldsymbol{\theta} \in \mathbb{R}^d$ if the characteristic function of its distribution can be expressed as: $\exp(i\boldsymbol{t}^T\boldsymbol{\theta})\phi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$, where $\boldsymbol{\Sigma}_{d\times d} \geq 0$ and ϕ are parameters. This definition can be found in an overview article [43]. More intuitive (but less general) definition of elliptical symmetry can be expressed under the assumption that the density of $\boldsymbol{X}$ exists. In this situation the density must be of the following form:

$$|\boldsymbol{\Sigma}|^{-1/2} g\left((\boldsymbol{X}-\boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}-\boldsymbol{\theta})\right)$$

  for some nonnegative function $g(\cdot)$ and some $\boldsymbol{\Sigma}_{d\times d} \geq 0$,
- *centrally symmetric* about $\boldsymbol{\theta} \in \mathbb{R}^d$ if $\mathcal{L}(\boldsymbol{X}-\theta) = \mathcal{L}(\boldsymbol{\theta}-\boldsymbol{X})$ where $\mathcal{L}(\cdot)$ denotes distribution of appropriate random vector; sometimes the term centrosymmetric is used instead of centrally symmetric,
- *angularly symmetric* about $\boldsymbol{\theta} \in \mathbb{R}^d$ if $\mathcal{L}((\boldsymbol{X}-\boldsymbol{\theta})/\|\boldsymbol{X}-\boldsymbol{\theta}\|) = \mathcal{L}((\boldsymbol{\theta}-\boldsymbol{X})/\|\boldsymbol{X}-\boldsymbol{\theta}\|)$, that is if the vector $(\boldsymbol{X}-\boldsymbol{\theta})/\|\boldsymbol{X}-\boldsymbol{\theta}\|$ is centrally symmetric. For absolutely continuous distributions the angularly symmetry coincides with the following notion of halfspace symmetry. It is easy to see that in such a case the point of symmetry is unique,
- *halfspace symmetric* about $\boldsymbol{\theta} \in \mathbb{R}^d$ if $\mathrm{P}(\boldsymbol{X} \in \mathbb{H}) \geq 1/2$ for every closed halfspace $\mathbb{H}$ containing $\boldsymbol{\theta}$.

In all previous cases $\boldsymbol{\theta} \in \mathbb{R}^d$ is called the centre of symmetry. It is easy to show that each of these notions of symmetry generalizes preceding one. Their relationship can be visualized by a simple scheme:

$$\text{elliptical sym.} \Rightarrow \text{central sym.} \Rightarrow \text{angular sym.} \Rightarrow \text{halfspace sym.}$$

The requirement of maximality at centre is very natural for distributions like multivariate normal distribution or uniform distribution on a convex support. However, there are also situations, in which the requirement is a bit problematic, for example uniform distribution on the support, which is the union of two triangles symmetric around their only common point. This point is the centre of symmetry, but apparently lies at the border of the support.

- *Monotonicity relative to the deepest point.* The depth of points $\boldsymbol{x} \in \mathbb{R}^d$ lying on some fixed ray going through the deepest point (point with the highest depth) should be a monotone (nonincreasing) function of the Euclidean distance of $\boldsymbol{x}$ from the deepest point. The property of monotonicity relative to the deepest point is at least ambiguous. Here we discuss two examples of datasets for which the property is not relevant. The first example can be described as random vector with uniform distribution on a "flower-shaped" support. Points randomly generated from this distribution are shown in Figure 1.1 on the left. Another example is the case of uniform distribution on a support, which consists of a unit circle and an annulus with the same center, see Figure 1.1 on the right.

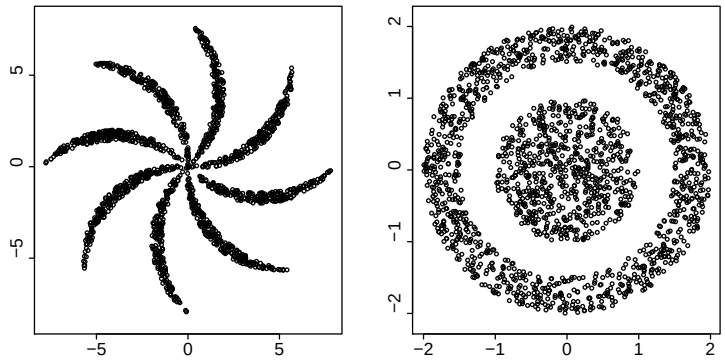


Figure 1.1: Randomly generated data from uniform distribution on a "flower-shaped" support (left) and on a "circle-annulus-shaped" support (right).

In both cases if the property of monotonicity relative to the deepest point was satisfied, there would be points out of the support of random vector (points with density equal to zero) that have higher depth then some points in the support of random vector. This fact is in contrast to the general idea of data depth, which was expressed by Zuo and Serfling as follows: "depth function is any function which provides central-outward ordering of vectors with respect to some probabilistic distribution P". In our case, we will probably get central-outward ordering, but the relationship to the probabilistic distribution is very problematic.

Although some of the previously stated properties are ambiguous, they are used to define the term "statistical depth function". Following formal definition was firstly published by Zuo and Serfling ([56]):

Definition 1.1 ***General definition of statistical depth function.*** *Denote by $\mathcal{P}$ the class of distributions on the Borel sets on $\mathbb{R}^d$ and by $P_{\boldsymbol{X}}$ the distribution of a given random vector $\boldsymbol{X}$. Let the mapping $D(\cdot;\cdot) : \mathbb{R}^d \times \mathcal{P} \to \mathbb{R}$ be bounded, nonnegative, and satisfy following conditions:*

1. *$D(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}; P_{\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}}) = D(\boldsymbol{x}; P_{\boldsymbol{X}})$ holds for any random vector $\boldsymbol{X}$ in $\mathbb{R}^d$, any $d \times d$ nonsingular matrix $\boldsymbol{A}$ and any d–dimensional vector $\boldsymbol{b}$;*
2. *$D(\boldsymbol{x}; P) \to 0$ as $\|\boldsymbol{x}\| \to \infty$ for each $P \in \mathcal{P}$;*
3. *$D(\boldsymbol{\theta}; P) = \sup_{\boldsymbol{x}\in\mathbb{R}^d} D(\boldsymbol{x}; P)$ holds for any $P \in \mathcal{P}$ having centre in $\boldsymbol{\theta}$;*
4. *$D(\boldsymbol{x}; P) \leq D(\boldsymbol{\theta} + \alpha(\boldsymbol{x} - \boldsymbol{\theta}), P)$ holds for $\alpha \in [0, 1]$, for any $P \in \mathcal{P}$ having the deepest point in $\boldsymbol{\theta}$.*

Then $D(\cdot; P)$ is called a statistical depth function.

In further text we use following notations and definitions of terms, which exhibit structure of multivariate distributions and reveal the shape of datasets:

Definition 1.2 • *The set $\left\{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}; P) \geq t\right\}$ is called the level set of depth t. Its border is known as the contour of depth t.*

- *The set* $\left\{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) \geq t\right\}$ *is called the level set of density* t *(*f *denotes density function).*
- *The* p*-th central region* C_p *is defined as the smallest region enclosed by depth contours to amass probability* p*, that is* $C_p = \bigcap_t \{R(t) : \mathrm{P}(R(t)) \geq p\}$*, where* $R(t) = \left\{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}; P) > t\right\}$*.*

Sometimes the term *central area* is used instead of the term *central region*.

1.2 Depth functions - an overview

In this section we would like to provide an overview of the most important and most widely used depth functions. We state their definitions and show their basic properties. Computational aspects are also discussed, because the fast computation of the empirical depth is crucial for relevant practical applications. Parts 1.2.1 and 1.2.2 are devoted to the halfspace depth and simplicial depth, that are probably the most often considered depth functions. The later two parts (1.2.3 and 1.2.4) are devoted to the L_1-depth and zonoid depth. These two notions of data depth do not satisfy the conditions of general depth function definition and hence they are not statistical depth functions in the sense of Definition 1.1. Nevertheless, they are used in applications because of their computational simplicity. The last part of this section summarizes other possible notions of data depth.

1.2.1 Halfspace depth

Definition 1.3 *The halfspace depth of a point* $\boldsymbol{x}$ *in* $\mathbb{R}^d$ *with respect to a probability measure* P *is defined as the minimum probability mass carried by any closed halfspace containing* $\boldsymbol{x}$*, that is*

$$D(\boldsymbol{x}; P) = \inf_{\mathbb{H}} \left\{\mathrm{P}(\mathbb{H}) : \mathbb{H} \text{ a closed halfspace in } \mathbb{R}^d : \boldsymbol{x} \in \mathbb{H}\right\}.$$

Sometimes it is useful to write the previous formula in the following form:

$$D(\boldsymbol{x}; P) = \inf_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \mathrm{P}\left(\left\{\boldsymbol{y} : \boldsymbol{u}^T(\boldsymbol{y} - \boldsymbol{x}) \geq 0\right\}\right).$$

The halfspace depth is sometimes also called location depth or Tukey depth as it was firstly defined by J. Tukey in 1975 ([52]). The halfspace depth is well defined for all $\boldsymbol{x} \in \mathbb{R}^d$. Its sample version (empirical halfspace depth), defined on a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the distribution P, is defined as a halfspace depth for the empirical probability measure P_n.

This definition is very intuitive and easily interpretable. Moreover, there are many nice properties of the halfspace depth, which made this depth function very popular and widely used. In particular, the halfspace depth function satisfies all desirable properties of Definition 1.1. This was proved by Zuo and Serfling in [56]. They classed the halfspace depth as a Type D depth function. This class of depth functions is defined as follows:

Definition 1.4 *Let* $\mathcal{C}$ *be a class of closed subsets of* $\mathbb{R}^d$ *and* P *a probability measure on* $\mathbb{R}^d$*. A corresponding Type D function is defined by*

$$D(\boldsymbol{x}; P, \mathcal{C}) = \inf_{C \in \mathcal{C}} \{\mathrm{P}(C) : \boldsymbol{x} \in C\}.$$

Some important properties (like upper semicontinuity or convex, compact and nested properties of central regions) are proved for the whole Type D class of depth functions under reasonable mild assumptions on $\mathcal{C}$:

Theorem 1.1 *Let $\mathcal{C}$ be a class of closed Borel sets satisfying both of following conditions:*

- *if $C \in \mathcal{C}$, then $\overline{C^c} \in \mathcal{C}$,*
- *for $C \in \mathcal{C}$ and $\boldsymbol{x} \in C^o$, there exist $C_1 \in \mathcal{C}$ with $\boldsymbol{x} \in \partial C_1, C_1 \subset C^o$,*

where $\partial C, C^c, C^o, \overline{C}$ denote, respectively, the boundary, complement, interior and closure of C,
Further, for a given probability measure P on $\mathbb{R}^d$, assume that if $\boldsymbol{x} \in C \in \mathcal{C}$ and $P(C) < p$, then there is a $C_1 \in \mathcal{C}$ such that $\boldsymbol{x} \in C_1^o$ and $P(C_1) < p$. Denote $R(p) = \left\{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}; P, \mathcal{C}) \geq p\right\}$, where $p \in (0, 1]$, regions with the highest depth. Then:

1. *$D(\boldsymbol{x}; P, \mathcal{C})$ is upper semicontinuous;*
2. *$R(p), p \in (0, 1]$ are compact and nested, i.e. $R(p_1) \subset R(p_2)$ if $p_1 > p_2$;*
3. *$R(p)$ is convex if every $C \in \mathcal{C}$ is convex.*

Proof: Can be found in [56] as the proof of Theorem 2.11.

Asymptotical properties of empirical halfspace depth were examined by Massé ([40]). He proved that under some mild conditions on P and fixed $\boldsymbol{x}$ such that $D(\boldsymbol{x}, P) = \alpha > 0$, it holds that $\sqrt{n}\,(D(\boldsymbol{x}, P_n) - D(\boldsymbol{x}, P))$ is asymptotically distributed as $N\,(0, \alpha(1 - \alpha))$.

Computational aspects of halfspace depth have been studied not only by statisticians, but also by the means of discrete geometry and optimization. A short summary of used approaches and the most important results can be found in [16]. Exact algorithms for computing halfspace depth in higher dimensions are of very limited use because of their high complexity. More precisely, the problem is known to be NP–complete (in meaning of complexity theory, see for example [17]) when number of points n and dimension d are both arbitrary. This was proved by Johnson and Preparata by showing the computation of halfspace depth to be equivalent to the *densest hemisphere problem* ([27]). This problem can be formulated as follows:

Let $\mathbb{R}^d$ be the d-dimensional Euclidean space and let $\mathbf{S}^d$ be the sphere of unit radius with centre at the origin of $\mathbb{R}^d$. Let K be a set of n points on $\mathbf{S}^d$. Find a hemisphere of $\mathbf{S}^d$ which contains a largest subset of K.

Johnson and Preparata also showed that the problem is equivalent to the problem of finding the *maximum feasible subsystem* of a system of strict homogeneous linear inequalities. This provides a different way for computing empirical halfspace depth. This approach led to so-called output-sensitive algorithms ([5]).

If the number d, which determines the dimension of Euclidean space, is fixed, then there exist an algorithm, which solves the problem in time $O(n^d)$. Johnson and Preparata proposed an enhanced algorithm, whose solving-time is $O(n^{d-1} \log n)$.

However, its use is very limited in higher dimensions (authors do not recommend use in dimension higher than $d = 4$). The idea of the naive recursive algorithm can be sketched as follows:

Lets consider the set $M = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of n nonzero points in $\mathbb{R}^d$. The algorithm works in n basic steps indexed by the letter i. Initialize $m := 0$. In the i-th step:

if $d = 1$ let $M_L = \{j : \boldsymbol{x}_j < \boldsymbol{x}_i\}$, $M_G = \{j : \boldsymbol{x}_j > \boldsymbol{x}_i\}$
and $m_{new} = \min(\mathrm{card}(M_L), \mathrm{card}(M_G))$.
If $m_{new} > m$ then redefine $m := m_{new}$
and $M^* := M_L$ (if $\mathrm{card}(M_G) < \mathrm{card}(M_L)$ then $M^* := M_G$).

if $d > 1$ define $H(\boldsymbol{x}_i) = \left\{\boldsymbol{y} \in \mathbb{R}^d : \boldsymbol{y}^T \boldsymbol{x}_i = 0\right\}$, the hyperplane of all vectors orthogonal to $\boldsymbol{x}_i$. Find the projection of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ into $H(\boldsymbol{x}_i)$:

$$\boldsymbol{x}'_j = \boldsymbol{x}_j - \frac{\boldsymbol{x}_j^T \boldsymbol{x}_j}{|\boldsymbol{x}_i|^2} \boldsymbol{x}_i \quad \text{for} \quad j = 1, \ldots, n$$

Solve the problem for nonzero points $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n$ in R^{d-1}.

The halfspace containing the largest number of points is determined as $M \setminus M^*$.

Most widely used algorithms for computing empirical halfspace depth are those by Rousseeuw et al. For instance, the package `depth`, which includes depth functions tools for multivariate data analysis in software R, uses exact algorithms published in [45] and [47] for dimensions two and three and approximate algorithms published in [47] for higher dimensions.

The latest development is focused on so called primal-dual algorithms - algorithms that update both upper and lower bounds of the depth. They rely on sufficiency of partial information and thus might terminate much earlier than exact algorithms. The algorithm has been developed mainly by Fukuda and Rosta and can be found in [4].

1.2.2 Simplicial depth

Definition 1.5 *The simplicial depth of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P is defined as*

$$D(\boldsymbol{x}; P) = \mathrm{P}(\boldsymbol{x} \in S[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}]),$$

where $S[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}]$ is a simplex with vertices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}$, that are independent identically distributed observations from P.

Simplicial depth was defined by R.Y. Liu in 1988 ([35]). The classical reference is her later paper [36] (dated 1990), where more detailed discussion of simplicial depth properties was presented. The simplicial depth function has all properties demanded by the Definition 1.1 providing that P is continuous angularly symmetric distribution (see [56]). However, Zuo and Serfling found some counterexamples showing that simplicial depth does not satisfy all conditions in general. They classed the simplicial depth as a Type A depth function. This class of depth functions is defined as follows:

Definition 1.6 *Let $h(\boldsymbol{x}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_r)$ be any bounded nonnegative function which in some sense measures the closeness of $\boldsymbol{x}$ to the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$. A corresponding Type A depth function is then defined by the mean closeness of $\boldsymbol{x}$ to the random sample of size r:*

$$D(\boldsymbol{x}; P) = \mathrm{E}_P\, h(\boldsymbol{x}; \boldsymbol{X}_1, \ldots, \boldsymbol{X}_r),$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r$ is a random sample from P.

In particular, we have $r = d+1$ and $h(\boldsymbol{x}; \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}) = I(\boldsymbol{x} \in S[\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_{d+1}}])$ for the simplicial depth.

Sample version of any Type A depth defined on a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of the distribution P is defined as the depth with respect to the empirical probability measure P_n:

$$D(\boldsymbol{x}; P_n) = \frac{1}{\binom{n}{r}} \sum_{*} h(\boldsymbol{x}; \boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_r}),$$

where $*$ denotes summing over all r-tuples $i_1, \ldots, i_r$ from the index set $1, \ldots, n$. Thus the sample version of any Type A depth has a form of U–statistic, as it was defined by Hoeffding in 1948 (see [23]). This knowledge enables use of general theory of U–statistics for examination of empirical Type A depth functions asymptotic properties.

Since $\mathrm{E}_P\, I(\boldsymbol{x} \in S[\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_r}]) = \mathrm{P}(\boldsymbol{x} \in S[\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_r}])$, we have the property of unbiasedness of empirical Type A depth function as it holds $\mathrm{E}\, D(\boldsymbol{x}; P_n) = D(\boldsymbol{x}; P)$ for any $\boldsymbol{x} \in \mathbb{R}^d$. It might be noted here that the letter "U" in the term "U–statistic" is right due to the property of unbiasedness. Nevertheless, more important property of U–statistics is its symmetry with respect to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Among others, Hoeffding derived formula for asymptotic variance of U–statistics: $\mathrm{Var}(U_n) \sim o(n^{-1})$. Hence we see that the convergence is as fast as for example the convergence of sample mean.

In currently existing R-package `depth`, the simplicial depth is computed in dimension 2 only ([41]). Calculation is exact and based on Fortran code by Rousseeuw and Ruts ([45]). This program reduces the number of operations to $O(n \log n)$ by combining some geometric properties with certain sorting and updating mechanism. In principle, in any finite dimension the simplicial depth could be calculated by determining whether or not a point is inside a random simplex. This can be done by checking if the point can be expressed as a convex combination of the vertices of the simplex. Such a naive algorithm needs to solve a system of linear equations. This naive routine is of course time–consuming, it takes $O(n^{d+1})$ time. Unfortunately, there does not exist any more effective algorithm for dimension higher than four so far. Cheng and Quyang gave an $O(n^2)$ algorithm for $\mathbb{R}^3$ and an $O(n^4)$ algorithm for $\mathbb{R}^4$ ([7]).

1.2.3 L_1–depth

Definition 1.7 *Let P be a probability distribution of a random vector $\boldsymbol{X}$ on $\mathbb{R}^d$ such that $\mathrm{E}_P \|\boldsymbol{X}\| < \infty$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$. The L_1–*

median of the distribution P is defined as

$$M(P) := \arg\min_{\boldsymbol{y}\in\mathbb{R}^d} \mathrm{E}_P \left\|\boldsymbol{X}-\boldsymbol{y}\right\|.$$

The L_1-depth of a point $\boldsymbol{x}\in\mathbb{R}^d$ with respect to the distribution P is defined as

$$D(\boldsymbol{x};P) = 1 - \inf\left\{w\geq 0 : M\left(\frac{w}{1+w}\delta_{\boldsymbol{x}} + \frac{1}{1+w}P\right) = \boldsymbol{x}\right\},$$

where $\delta_{\boldsymbol{x}}$ is a point mass at $\boldsymbol{x}$. That is, $1-D(\boldsymbol{x};P)$ is the minimum incremental mass w needed at $\boldsymbol{x}$ for $\boldsymbol{x}$ to become the L_1–median of the mixture $\frac{w}{1+w}\delta_{\boldsymbol{x}} + \frac{1}{1+w}P$.

The L_1–depth is also sometimes called the spatial depth. The definition of L_1–depth was proposed by Vardi and Zhang in 2000 ([54]). They provided an interesting historical overview of problems which motivated definition of L_1–median, namely Fermat–Weber location problem (a logistic problem of optimal selection of a location). They also provided a modification of Weiszfeld's iterative algorithm for computing empirical L_1–median and emphasized high breakdown point (equal to 1/2) of L_1–median. Probably the most important contribution of the cited work is the derivation of a formula for empirical L_1–depth, which enables its fast computation:

Theorem 1.2 *Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ be a random sample from absolutely continuous distribution P on $\mathbb{R}^d$ such that $\mathrm{E}_P\left\|\boldsymbol{X}\right\|<\infty$.*

- *For $\boldsymbol{x}\notin\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\}$ denote $\bar{e}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \frac{\boldsymbol{x}-\boldsymbol{X}_i}{\|\boldsymbol{x}-\boldsymbol{X}_i\|}$ the spatial rank function. Then $D(\boldsymbol{x};P_n) = 1-\|\bar{e}(\boldsymbol{x})\|$.*
- *For $\boldsymbol{x}=\boldsymbol{X}_k$, for some $k\in\{1,\ldots,n\}$ denote $\bar{e}(\boldsymbol{x}) = \frac{1}{n-1}\sum_{i\neq k} \frac{\boldsymbol{x}-\boldsymbol{X}_i}{\|\boldsymbol{x}-\boldsymbol{X}_i\|}$ the spatial rank function. Then $D(\boldsymbol{x};P_n) = 1-(\|\bar{e}(\boldsymbol{x})\|-1/n)^+$.*

Proof: Derivation can be found in Section 4 of [54].

Note that when computing the depth of some sample point $\boldsymbol{X}_k, k\in 1,\ldots,n$ the presence of the point in sample "increases" its depth. The theorem above is stated in a simple version under the assumption that there are no ties in the data. This assumption is quite natural when considering random sample from absolutely continuous distribution, but it is not sometimes satisfied in practice. More general version can be found in [54]. Other works devoted to L_1–depth sometimes use more straight definition than Definition 1.7; for example Serfling ([49]) defines L_1–depth of the point $\boldsymbol{x}\in\mathbb{R}^d$ with respect to the distribution P by the formula:

$$D(\boldsymbol{x};P) = 1 - \left\|\mathrm{E}_P\left(\frac{\boldsymbol{x}-\boldsymbol{X}}{\|\boldsymbol{x}-\boldsymbol{X}\|}\right)\right\|.$$

He also recalls some older results to show consistency of the empirical L_1–depth.

Although Vardi and Zhang showed some good properties of L_1–depth (for example vanishing at infinity), the L_1–depth does not satisfy the property of affine invariance and hence it is not a statistical depth function in a sense of Definition 1.1.

Computational simplicity can be considered as one of the biggest advantages of L_1–depth. It makes it attractive for applications in high dimensional spaces. Counting the depth is easy and hence there is no need of preprogrammed routines. Estimating L_1–median, which is a more complicated problem, is implemented in (at least) two R–packages: `ICSNP` (which deals with tools for multivariate non-parametrics) uses directly algorithm by Vardi and Zhang, and `pcaPP` (which deals with robust principal component analysis by projection pursuit).

1.2.4 Zonoid depth

The zonoid depth was proposed by Koshevoy and Mosler at mid ninetieth of the 20th century. The definition of zonoid region itself demands our attention:

Definition 1.8 ***Zonoid regions:*** *The α–zonoid region of some distribution P on $\mathbb{R}^d$ with the finite first moment is defined as*

$$Z_\alpha(P) = \left\{ \int_{\mathbb{R}^d} \boldsymbol{x} g(\boldsymbol{x}) dP(\boldsymbol{x}) : \ g : \mathbb{R}^d \to [0, 1/\alpha] \, , \int_{\mathbb{R}^d} g(\boldsymbol{x}) dP(\boldsymbol{x}) = 1 \right\}$$

for $\alpha \in (0, 1]$ and $Z_0(P) = \mathbb{R}^p$.

Obviously $\int_{\mathbb{R}^d} \boldsymbol{x} g(\boldsymbol{x}) dP(\boldsymbol{x})$ is a point in $\mathbb{R}^d$, hence $Z_\alpha(P)$ is a set of points in $\mathbb{R}^d$ which rises by successive choosing of all possible "weight" functions $g(\cdot)$. This functions are requested to have two above mentioned properties: $g(\boldsymbol{x}) \in [0, 1/\alpha] \ \forall \boldsymbol{x} \in \mathbb{R}^d$ and $\int_{\mathbb{R}^d} g(\boldsymbol{x}) dP(\boldsymbol{x}) = 1$ (note the similarity to the property of a density function).
When considering empirical distribution function with equal probability mass $1/n$ at each point $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and denoting $g(\boldsymbol{X}_i) = \lambda_i^*$ for $i = 1, \ldots, n$, the previous formula can be written as:

$$Z_\alpha(P_n) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \lambda_i^* \boldsymbol{X}_i : \frac{1}{n} \sum_{i=1}^{n} \lambda_i^* = 1, 0 \leq \lambda_i^* \leq 1/\alpha \text{ for all } i \right\}.$$

It is easy to see that:

- zonoid regions are nested, that is if $\alpha_1 \geq \alpha_2$ then $D_{\alpha_1} \subset D_{\alpha_2}$,
- $Z_1(P) = \mathrm{E}_P \boldsymbol{X}$,
- zonoid regions for $\alpha > 0$ are affine equivariant, bounded and convex.

Definition 1.9 ***Zonoid depth:*** *The zonoid depth of a point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to distribution P on $\mathbb{R}^d$ is defined as*

$$D(\boldsymbol{x}; P) = \max \{ \alpha : \boldsymbol{x} \in Z_\alpha(P) \} .$$

Theoretical properties and possibilities of generalization are studied for example in [31] and [32]. It is proved that the zonoid depth is affine invariant and vanishes at infinity. Further it is proved that the zonoid depth is upper semi-continuous (the set $\left\{ \boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}; P) \geq \alpha \right\}$ is closed for every α) and quasi-concave (for any $\lambda \in [0, 1]$ and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$ it holds $D(\lambda \boldsymbol{x}_1 + (1 - \lambda) \boldsymbol{x}_2; P) \leq \max \{ D(\boldsymbol{x}_1; P), D(\boldsymbol{x}_2; P) \}$). However, it can fail to satisfy property of maximality

at centre when the distribution is not centrally symmetric, because it attains its maximal value always at $\mathrm{E}_P \boldsymbol{X}$. This also brings a usual problem of nonrobustness of sample zonoid depth.

Computational aspects of zonoid depth are discussed in detail in [12]. It was shown that the computation of zonoid depth of a point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to empirical distribution P_n based on random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ can be done by solving the following linear program:

$$\begin{aligned} &\text{Minimize } \gamma \\ &\text{subject to:} \quad \boldsymbol{X}\boldsymbol{\lambda} = \boldsymbol{x} \\ &\qquad\qquad\quad \boldsymbol{\lambda}\mathbf{1} = 1 \\ &\qquad\qquad\quad \gamma\mathbf{1} - \boldsymbol{\lambda} \geq \mathbf{0}, \quad \boldsymbol{\lambda} \geq \mathbf{0} \end{aligned}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is the data matrix whose columns are vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $\mathbf{0}$ and $\mathbf{1}$ denote n-dimensional (column) vectors of zeros, ones respectively. $\boldsymbol{\lambda} = 1/n(\lambda_1^*, \ldots, \lambda_n^*)^T$ is a vector of unknown real parameters. If γ^* is the optimal value of the objective, then it holds:

$$D(\boldsymbol{x}; P_n) = \frac{1}{n\gamma^*}.$$

Koshevoy et el. do not recommend standard simplex methods for solving this linear program when n is too large. They provided an algorithm which takes advantage of the special structure of the set of constraints by a Dantzig–Wolfe decomposition. This very fast algorithm for computing the zonoid depth makes this depth very attractive for use in applications (see for example [42], where it is used for classification problem). The algorithm can be found in [12]. Current computer can count zonoid depth of 1000 points in 8-dimensional space in few minutes. The same computation of the (exact) halfspace depth would exceed few years.

1.2.5 Other depth functions

Except the previously mentioned depth functions, many other notions of data depth have been introduced so far. We do not describe them in detail here. Nevertheless, they should be mentioned at least. We do not give reference to particular depth functions. They can be found in [38], [50] or [56].

- The **Mahalanobis depth** was introduced by Mahalanobis in 1936. It is defined as
$$D(\boldsymbol{x}; P) = \frac{1}{1 + (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})},$$
where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the dispersion matrix of P. The definition is quite easy to understand and the computation of empirical Mahalanobis depth is quite fast. The problem is that the depth is very closely related to elliptically symmetric distributions - the levelsets of the depth are ellipsoids for any P, even if P is not elliptically symmetric. Moreover, there is no unique way how to estimate parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, classical estimators

are not robust. The depth is nonzero on the whole space, what can be useful in some applications. Zuo and Serfling placed this depth to the class of Type C depth functions ([56]).

- The **projection depth** was firstly defined by Liu in 1992 and inspected later mainly by Zuo. It was inspired by the projection idea behind the Stahel–Donoho estimator. The projection depth of the point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to distribution P is defined as:

$$D(\boldsymbol{x}; P) = \frac{1}{1 + O(\boldsymbol{x}; P)}, \quad \text{where} \quad O(\boldsymbol{x}; P) = \sup_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \frac{\left|\boldsymbol{u}^T\boldsymbol{x} - \text{med}(\boldsymbol{u}^T\boldsymbol{X})\right|}{\text{MAD}(\boldsymbol{u}^T\boldsymbol{X})},$$

  where random vector $\boldsymbol{X}$ has distribution P, $\text{med}(Y)$ denotes median and $\text{MAD}(Y) = \text{med}\left(|Y - \text{med}(Y)|\right)$ denotes the univariate median absolute deviation. More generally, some other measures of location and dispersion could be considered instead of the median and MAD. Notice that the formula for projection depth has similar structure as the formula for Mahalanobis depth. The projection depth belongs to the same group of Type C depth functions according to [56]. It is a statistical depth function in the sense of Definition 1.1 (it has all four desirable properties). This makes it very attractive. However, quite a small attention has been paid to this notion of data depth so far.

- The **convex hull peeling depth** was introduced by Barnett in 1976. It is defined for some sample point $\boldsymbol{X}_k,\ k = 1, \ldots, n$ with respect to the data set $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ as the level of the convex layer $\boldsymbol{X}_k$ belongs to. The computation is quite fast, but there is no theoretical version (population analogue) of this sample depth. Donoho and Gasko discussed breakdown properties of the "peeled mean" - mean of the points with the highest convex hull peeling depth - in [10]. They showed (in section 4.1) that the breakdown point can not be higher than $1/(d+1)$. Consequently, they placed this method among those which do not attain high breakdown point.

- The **likelihood depth** was introduced by Fraiman and Meloche in 1996. It is defined as density, that is

$$D(\boldsymbol{x}; P) = f(\boldsymbol{x}),$$

  where f is density function of underlying distribution. This definition brings the problem of density estimation with all its difficulties. This notion of depth does not generally satisfy any of the four basic properties of Definition 1.1. This notion of depth is particularly useful when describing multimodality of a distribution.

- The **Oja depth** and the **simplicial volume depth** are both based on the volume of random simplex. The later (and more complicated) one takes into consideration a dispersion of the distribution.
  The Oja depth of a point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to distribution P is defined as:

$$D(\boldsymbol{x}; P) = \left[1 + \mathrm{E}_P\, \Delta(S[\boldsymbol{x}; \boldsymbol{X}_1, \ldots, \boldsymbol{X}_d])\right]^{-1}.$$

The simplicial volume depth of a point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to distribution P is defined as:

$$D(\boldsymbol{x}; P) = \left[1 + \mathrm{E}_P \left(\frac{\Delta(S[\boldsymbol{x}; \boldsymbol{X}_1, \ldots, \boldsymbol{X}_d])}{\sqrt{\det(\boldsymbol{\Sigma})}} \right)^{\alpha} \right]^{-1},$$

where $\Delta(S[\boldsymbol{x}; \boldsymbol{X}_1, \ldots, \boldsymbol{X}_d])$ denotes the volume of d–dimensional closed simplex $S[\boldsymbol{x}; \boldsymbol{X}_1, \ldots, \boldsymbol{X}_d]$ with vertices $\boldsymbol{x}$ and d random observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d$ from P, $\boldsymbol{\Sigma}$ is the covariance matrix of P and $\alpha > 0$ is a parameter. The sample versions are analogous to the sample version of the simplicial depth. Oja depth is not affine invariant in general. Zuo and Serfling proved ([56]) that the simplicial volume depth is a depth function in the sense of Definition 1.1 under the assumption that $\alpha \geq 1$ and P being centrally symmetric. They classified the depth as Type B depth function.

- The $\boldsymbol{L^p}$**–depth** $\boldsymbol{(p > 0)}$ functions use the L^p norm $\|\cdot\|_p$ in definition. The L^p–depth of a point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to distribution P is defined as

$$D(\boldsymbol{x}; P) = \left(1 + \mathrm{E}_P \|\boldsymbol{x} - \boldsymbol{X}\|_p\right)^{-1}.$$

  Notice that the formula in definition is similar to the formula for simplicial volume depth. The L_p–depth functions are also classified as Type B depth functions by Zuo and Serfling. The affine invariant version of L^2–depth is particularly interesting. It satisfies all desirable properties of Definition 1.1 under the assumption of P being angularly symmetric about a unique point.

- The **majority depth** was proposed by Singh in 1991. The majority depth of a point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to P is defined as

$$D(\boldsymbol{x}; P) = P\left(\boldsymbol{x} \in M_P(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)\right),$$

  where $M_P(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$ denotes so called major side determined by $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d$, that is the halfspace bounded by hyperplane containing $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d$ which has probability $\geq 1/2$. One can see from Definition 1.6 that the majority depth is a Type A depth function with $r = d$ and $h(\boldsymbol{x}; \boldsymbol{X}_1, \ldots, \boldsymbol{X}_d) = I\left(\boldsymbol{x} \in M_P(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)\right)$. The majority depth satisfies properties of maximality at centre and monotonicity relative to deepest point under the assumption of P being halfspace symmetric.

- Serfling ([50]) got together several *extended notions of depth function*. Probably the most well known is the **regression depth**, which was introduced by Rousseeuw and Hubert in 1999. It provides an extension of halfspace and simplicial depth and defines depth notion in the regression settings. In this approach, the depth is a property of a fit (represented by coefficients $\boldsymbol{\beta}$). The depth of a fit $\boldsymbol{\beta}$ with respect to empirical distribution P_n determined by random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is defined as the smallest number of observations that would need to be removed in order to make $\boldsymbol{\beta}$ a nonfit. Nonfits are exactly those fits with zero depth. A particular definition of a nonfit determines a particular depth function. For more details see [44]. Another generalizations of depth function: **data depth on circles and spheres**

by Liu and Singh (1992), **tangent depth** by Mizera (2002), **generalized forms of Tukey depth** by Zhang (2002) and **location–scale depth** by Mizera and Müller (2004).

- New approaches to data depth have come from discrete geometry, for example the class of **proximity graph data measures** such as Delaunay depth, Gabriel graph depth or β–skeleton depth introduced by Rafalin, Seyboth and Souvaine or the **colourful simplicial depth** by Deza et al. For detailed reference see [26].

1.3 Possible applications of the data depth

This section provides author's selection of some interesting applications of data depth methodology published in literature. From obvious reasons, the list of applications is not exhaustive.

The notion of *rank* is crucial in many applications. Consider a d-dimensional probability distribution P and a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from this distribution. (The empirical probability measure based on the sample is denoted by P_n). For any point $\boldsymbol{x} \in \mathbb{R}^d$ we define

$$r_P(\boldsymbol{x}) = \mathrm{P}(D(\boldsymbol{X}; P) \leq D(\boldsymbol{x}; P) \mid \boldsymbol{X} \sim P) \tag{1.1}$$

and

$$r_{P_n}(\boldsymbol{x}) = \#\left\{\boldsymbol{X}_i : D(\boldsymbol{X}_i; P_n) \leq D(\boldsymbol{x}; P_n),\ i = 1, \ldots, n\right\}/n. \tag{1.2}$$

- **Outlier detection - a bagplot**. Rousseeuw et al. [46] proposed a bivariate generalization of the univariate boxplot, so called bagplot. They used the halfspace depth to order the data, but other depth functions might be used as well. The bagplot consists of
  - the deepest point (the point with maximal depth),
  - the bag, that is the central area, which contains 50% of all points; the bag is usually dark colored,
  - the fence, which is found by magnifying the bag by a factor 3; the fence is usually not plotted; observations outside the fence are flagged as outliers,
  - the loop, which is an area between the bag and the fence; usually light coloured.

  The bagplot procedure is available in `R` library `aplpack`. As an example, we used car data of Chambers and Hastie that are available in library `rpart`. Figure 1.2 displays car weight and engine displacement of 60 cars. Five outliers were detected.

- **Affine equivariant and robust estimates of location.** Donoho and Gasko ([10]) have shown that two basic location estimators based on the halfspace depth, the deepest point and the trimmed mean (with trimming based on the halfspace depth), are both affine equivariant and robust (in the

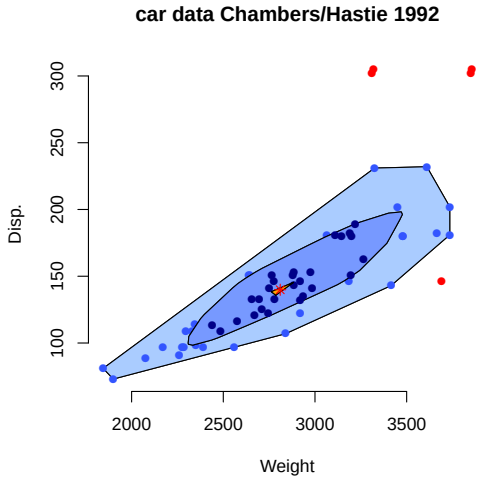


Figure 1.2: An example of bagplot.

sense of the high breakdown point). The combination of these two properties is quite rare in multivariate statistics. The most important results are summarized in the next theorem:

Theorem 1.3 *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a sample determining empirical version P_n of an absolutely continuous distribution P on $\mathbb{R}^d$, with $d > 2$. Assume data be in a general position (no ties, no more than two points on any line, three in any plane, and so forth).*
Consider the deepest point $T_*(P_n) = \arg\max_{\boldsymbol{x}} D(\boldsymbol{x}, P_n)$ and α-trimmed mean $T_\alpha(P_n) = Ave(\boldsymbol{X}_i : D(\boldsymbol{X}_i; P_n) \geq n\alpha)$, the average of all points whose depth is at least $n\alpha$.
Denote $\beta := \arg\max_{\boldsymbol{x}} D(\boldsymbol{x}; P)$ ($\beta = 1/2$ if P is centrally symmetric). Then

1. *The breakdown point of $T_*(P_n)$ is greater or equal to $1/(d+1)$. It converges almost surely to $1/3$ as $n \to \infty$ if P is centrally symmetric.*
2. *For each $\alpha \leq \beta/(1+\beta)$, $T_\alpha(P_n)$ is well defined for sufficiently large n and its breakdown point converges almost surely to α.*

Proof: Can be found in [10] as the proof of Propositions 3.2 - 3.4.

- **Rank tests for multivariate scale difference.** Liu and Singh [39] combined ranks based on data depth with well-known one-dimensional nonparametric procedures to test scale difference between two or more distributions.

  Consider two d-dimensional distributions P_1 and P_2, which possibly differ in dispersion only. Denote $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ a random sample from P_1 and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$ a random sample from P_2. Denote the combined sample as $\{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{n_1+n_2}\} \equiv \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}\}$ and denote $P_{n_1+n_2}$ the empirical distribution function based on the combined sample.

  We want to test the hypothesis H_0 of equal scales against the alternative that P_2 has larger scale in the sense that the scale of P_2 is an expansion of the

scale of P_1. If the scale of P_2 is greater, then obviously observations from the second distribution tend to be more outlying than the observations from P_1. Consider the sum of the non-normalized ranks for the sample from P_2:

$$R\left(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_{n_2}\right)=(n_1+n_2)\sum_{i=1}^{n_2} r_{P_{n_1+n_2}}(\boldsymbol{Y}_i).$$

Now we proceed as in the case of testing for a (negative) location shift in the univariate setting. This leads us to the Wilcoxon rank-sum procedure. When n_1 and n_2 are sufficiently large, we can rely on asymptotic behaviour of the test statistic (assuming null hypothesis):

$$R^*=\frac{R(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_{n_2})-[n_2(n_1+n_2+1)/2]}{[n_1n_2(n_1+n_2+1)/12]^{1/2}}\xrightarrow{D}N(0,1)$$

and hence we reject H_0 if $R^*\leq\Phi^{-1}(\alpha)$, where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard normal distribution.

We can proceed similarly when considering more than two (say $K>2$) distributions. We test the hypothesis that the underlying distributions are identical against the alternative that the scales of these distributions are not all the same, in the sense of scale contraction. Construction of the test follows the idea of the well-known Kruskal-Wallis test. Let $\bar{R}_i$ denote the average rank (based on data depth) of the observations from the i-th sample in the combined sample. The total number of all observations in combined sample (from all K samples) is N. Under the null hypothesis, it holds:

$$T=\frac{12}{N(N+1)}\sum_{i=1}^{K}\left(n_i\bar{R}_i^2\right)-3(N+1)\xrightarrow{D}\chi^2_{K-1}$$

We reject the null hypothesis at an approximate level α if $T\geq\chi^2_{K-1}(1-\alpha)$, where $\chi^2_{K-1}(1-\alpha)$ is the $(1-\alpha)$ quantile of a chi-squared distribution with $(K-1)$ degrees of freedom.

There is a simple graphical tool developed by Liu, Parelius and Singh (see [38]) to visualize difference in scales of multivariate distributions. They defined a *scale curve* as a plot of $p\in(0,1)$ versus volume of C_p - the p-th central region (see Definition 1.2). The sample scale curve, based on random sample $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$, plots volumes of the convex hulls containing $\lceil np\rceil$ most central points (versus p). By plotting scale curves for compared distributions in one plot, the difference in scales can be easily visualized.

The following example should illustrate the methodology. We simulated 250 points from bivariate $N(\mathbf{0},\boldsymbol{I})$ distribution and the same number of points from $N(\mathbf{0},2\boldsymbol{I})$ ($\boldsymbol{I}$ denoting 2×2 identity matrix). The test statistic was $R^*=-2,14$, which is less than $\Phi^{-1}(0,05)=-1,64$. We thus (correctly) reject the null hypothesis of identical distributions. The difference in dispersions can be seen in Figure 1.3.

- **Control charts for multivariate processes.** Liu [37] used the concept of data depth to introduce control charts for monitoring processes of multivariate quality measurements. The idea is to work with ranks of the multivariate

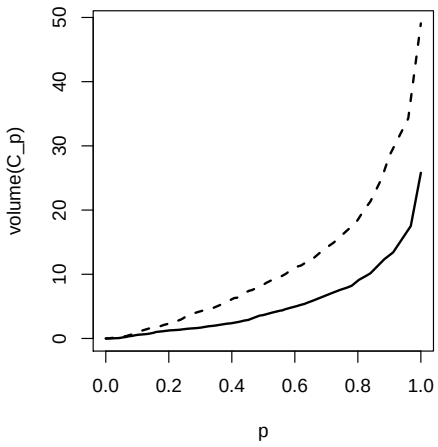


Figure 1.3: *Empirical scale curves based on samples of 250 points from* $N(\mathbf{0}, \boldsymbol{I})$ *(solid line) and from* $N(\mathbf{0}, 2\boldsymbol{I})$ *(dashed line).*

measurements (based on data depth) rather than with multivariate measurements themselves.

Let G denote the prescribed d-dimensional distribution (if the measurements follow the distribution G, the process is considered to be in control). G is either known or it can be estimated: G_n denotes its empirical version, based on n observations. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be the new observations from the considered process. They follow some distribution F. Our task is to test the null hypothesis $H_0 : F \equiv G$ against the alternative H_A: there is a location shift or a scale increase from G to F.

The test is based on ranks $r_G(\boldsymbol{X}_1), r_G(\boldsymbol{X}_2) \ldots$ (or $r_{G_n}(\boldsymbol{X}_1), r_{G_n}(\boldsymbol{X}_2), \ldots$ if G needs to be estimated). Under the null hypothesis, it holds:

1. $r_G(\boldsymbol{X}) \sim U[0, 1]$,
2. $r_{G_n}(\boldsymbol{X}) \xrightarrow{D} U[0, 1]$, provided that $D(\cdot; G_n) \to D(\cdot; G)$ uniformly as $n \to \infty$.

The uniform convergence of $D(\cdot; G_n)$ holds for example for simplicial depth if G is absolutely continuous. The expected value of $r_G(\boldsymbol{X})$ is thus 0.5. Small values correspond to a change in the process. A so-called lower control limit is thus equal to α (typically 0.05). Values $r_G(\boldsymbol{X}_i) < \alpha$ signalize a possible quality deterioration.

Similarly as Liu [37], we can demonstrate the procedure on simulated data. Let the prescribed distribution G be a bivariate standard normal distribution. Firstly, we generate 500 observations from this distribution to get a sample version G_n (we consider G to be unknown to mimic some real applications). Subsequently, we generate new observations - 40 observations from bivariate standard normal distribution (process in control) and next 40 observations from bivariate normal distribution with shifted mean $(2, 2)^T$ and both scales doubled. The control chart is shown in Figure 1.4.

There is one so called false alarm in the first half of observations. The out-of-control status in the second half of observations is correctly detected 30

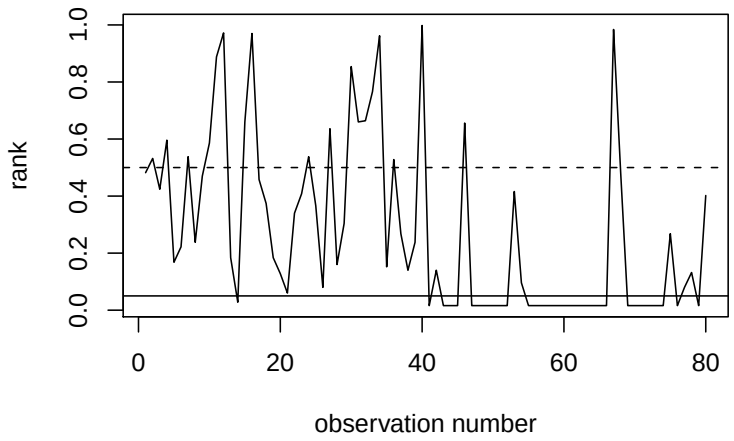


Figure 1.4: Control chart for multivariate process.

times (from 40 observations). The change is apparent from the chart.

Liu called this type of control chart the r chart. She also proposed multivariate versions of Shewhart chart (Q chart) and CUSUM chart (S chart).

- **Multivariate density estimation.** Fraiman, Liu and Meloche used the data depth for smart estimating of multivariate density function [15]. They considered the case in which a density function f of some multivariate probability distribution can be expressed as some real function of the data depth:

$$f(\boldsymbol{x}) = g\left(D(\boldsymbol{x}; P)\right).$$

The basic idea is to estimate data depth and subsequently to estimate the density function f as a function of D by a one-dimensional kernel density estimation. The advantage is that most of depth functions can be estimated at the usual parametric rate $1/\sqrt{n}$ and it does not affect the overall rate of convergence of the estimator. For the considered class of densities we get a one-dimensional nonparametric rate of convergence in a d-dimensional density estimation problem - the estimator based on data depth has much better asymptotic performance than the usual kernel density estimator. The assumption of the relationship between the depth and the density is crucial - the proposed estimates are not universally consistent.

- **Robust estimation of hydrological model parameters** was proposed by Bárdossy and Singh [2]. Hydrological models are used for different purposes such as water management or flood forecasting. The models have usually about ten parameters, that need to be estimated. Estimation routines are based on maximizing some objective function measuring model performance, for example so called Nash-Sutcliffe coefficient. Unfortunately, many of the hydrological observations that are used as inputs to estimation procedure contain partly systematic and partly random errors. This can strongly influence the model performance - the parameters obtained by sophisticated optimization procedures might be suboptimal in reality. Bárdossy and Singh investigated the set of parameters, which gives similar performance as the numerical optimum (so called good parameters). They proposed an iterative algorithm to find a convex set containing good model parameters:

1. Limits for the d selected parameters are identified.
2. N random parameter vectors forming the set S_0 are generated in the d dimensional rectangle bounded by the limits defined in 1.
3. Hydrological model is run for each parameter vector in S_0 and the corresponding model performances are calculated.
4. The subset S_0^* of the best performing parameters is identified. This might be for example the best 10% of S_0.
5. M random parameter sets forming the set S_1 are generated, such that for each parameter vector $\boldsymbol{\beta} \in S_1$, $D(\boldsymbol{\beta}) \geq L$, where the halfspace depth is calculated with respect to the set S_N^*.
6. The set S_1 is relabeled as S_0 and steps 3–6 are repeated until the performance corresponding to S_0 and S_1 does not differ more than what one would expect from the observation errors.

Authors argued that parameters with low data depth are near the boundary and are sensitive to small changes and do transfer to other time periods less well as high depth ones.

- **Clustering and classification - an analysis of microarray gene expression data**. Jörnsten applied the methodology of data depth in the analysis of several datasets that includes so called gene expression data [28]. A particularly important task is clustering of genes based on their expression levels under various experimental conditions.

  Jörnsten proposed a clustering algorithm (DDclust) which proceeds from some initial partition and improves it in some sense. The PAM (Partition Around Medoids) algorithm is used for initial partition. The DDclust algorithm computes so called relative data depth, which can be considered as a measure of cluster affiliation uncertainty, for all observations. Subsequently a random subset of points with low relative data depth (under a certain threshold) is relabeled. The relative data depth of an observation $\boldsymbol{x}_i$ is defined as:

  $$ReD_i = D(\boldsymbol{x}_i; P_k) - \max_{l \neq k} D(\boldsymbol{x}_i; P_l),$$

  where P_k is an empirical distribution of the points currently assigned to the same cluster as $\boldsymbol{x}_i$. The L_1-depth is used because of its computational simplicity and property of non-zero values outside of the convex hull of the data.

  Jörnsten concludes, that "A cross validation study on real gene expression data showed that DDclust was robust, and generated clusters could be used to predict sample labels better than PAM. Gene clustering with DDclust performed better than PAM in terms of the gene clusters' sample predictive properties."

1.4 Data depth - a useful/useless methodology

The concept of data depth has brought many questions and challenges. Originally, the data depth was proposed as a measure of outlyingness. The view of depth functions was opened up as a general nonparametric approach in late 80s mainly by Regina Liu. This broader view of data depth has brought questions like: *Can the depth function uniquely determine a probability distribution? Which well-known univariate concepts can be extended to the multivariate setting via depth functions? What are the strengths and weaknesses of the data depth concept?* We deal with these questions in this section.

First, the relation between the depth function and the probability measure should be clarified. The Cramér-Wold theorem (firstly published in 1936 [9]) states that a Borel probability measure on $\mathbb{R}^d$ is uniquely determined by all quantiles of all its one-dimensional projections. There is a natural question if the depth function could also uniquely determined a probability distribution. Koshevoy [30] has shown such a property for the halfspace depth and atomic measures with finite support:

Theorem 1.4 *Consider an atomic measure μ, which assigns masses $\mu_i > 0$ to points $\boldsymbol{x}_i$ of some dataset X. Denote $\mu = (M, X)$, where $M = \{\mu_1, \ldots, \mu_n\}$ and $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. If $\mu = (M, X)$ and $\nu = (M', X')$ are atomic measures such that, for any $\boldsymbol{x} \in \mathbb{R}^d$, $D(\boldsymbol{x}; (M, X)) = D(\boldsymbol{x}; (M', X'))$, then $\mu = \nu$.*

Hassairi and Regaieg [21] considered the case of absolutely continuous distributions with connected support:

Theorem 1.5 *Let μ be an absolutely continuous probability measure on $\mathbb{R}^d$ with connected support. Suppose that the probability density f of μ has second partial derivatives. Then the halfspace depth function of μ characterizes μ.*

The theorem was proved in slightly more general form. The condition of the second derivatives existence is not necessary.

Inspection of the data depth in one-dimensional case can clarify the role of data depth in multivariate case. Consider an absolutely continuous random variable X. Denote its distribution as P. Then (from Definition 1.3) the halfspace depth of any $x \in \mathbb{R}$ is defined as

$$D(x; P) = \min \{P(X \leq x), P(X \geq x)\}.$$

Obviously, the deepest point is median of P, for which the halfspace depth is equal to 1/2. The deepest point thus has all advantages and disadvantages of median in univariate case. Recall that for

- symmetric, unimodal distributions: $\mathrm{med}(X) = \mathrm{E}X = \mathrm{mod}(X)$,
- symmetric distributions: $\mathrm{med}(X) = \mathrm{E}X$.

Recall that median do not have to be unique. This could be illustrated by the well-known example of a mixture of two "equally probable" distributions with separated supports. The Figure 1.5 shows the density function of such a mixture.

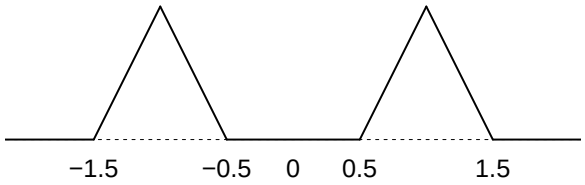


Figure 1.5: The median need not to be unique.

All points x in interval $[-0.5, 0.5]$ satisfy $P(X \leq x) \geq 0.5$ and $P(X \geq x) \geq 0.5$, which is the property that determines the median.

This example also emphasizes the importance of unimodality assumption. Without this assumption the most central points could lie in the area of small density, what could be undesirable under certain circumstances.

As the deepest point is the median in one dimensional case, it can be considered to be a generalization of median in multidimensional case. There are two different ways, in which the notion of median is usually extended to the multivariate setting:

- coordinate-wise median: $\text{med}(\boldsymbol{X}) = (\text{med}(X_1), \ldots, \text{med}(X_d))^T$.
- geometric median: $\text{med}(\boldsymbol{X}) = \arg\min_{\boldsymbol{y} \in \mathbb{R}^d} \mathrm{E}\, \|\boldsymbol{X} - \boldsymbol{y}\|$, which is also known as spatial median or L_1–median.

The coordinate-wise median is quite simple to compute, however it is not affine equivariant. It is even not rotation equivariant as can be seen from the Figure 1.6. In this example we consider uniform distribution on the "L"-shaped area, which is a union of rectangles $[0, 4] \times [0, 1]$ and $[0, 1] \times [0, 4]$. The component-wise median is the point $[7/8, 7/8]$ (Figure 1.6, left). After the $\pi/4$ rotation about the origin, we get the component-wise median $[0, 11\sqrt{2}/8]$, which is clearly not the image of the original component-wise median.

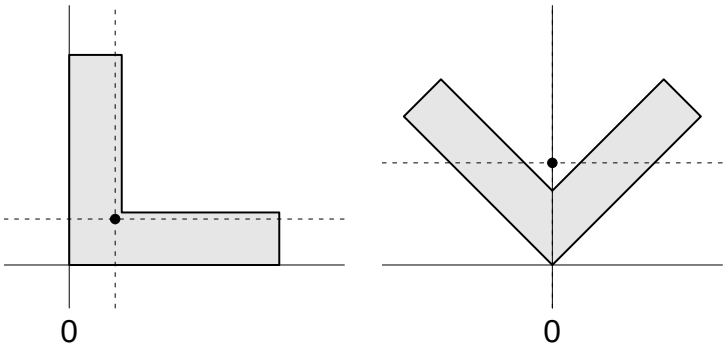


Figure 1.6: Component-wise median is not affine equivariant.

The geometric median is the deepest point when considering L_1-depth. This depth function is unfortunately not affine invariant and hence the median is not affine equivariant. We can obtain affine equivariant multivariate median when we define it as the deepest point of a distribution when using any affine invariant depth function, for example halfspace depth.

The example above shows that the concept of data depth could be used in a favourable way for extension of some univariate concepts to multidimensional setting. The data depth is useful mainly for generalizing methods based on ranks, quantiles and outlyingness measure. However, the depth function generally does not characterize the probability function - it is a simplification connected with reduction of information. The possibility of its use is limited, as was shown in the example of multimodal distribution.

Use of the data depth is meaningful when the dimension of the space is not too high. In our opinion, it should not be used when the dimension is higher than 10. This advice is given not only because of computational costs, but mainly for theoretical reasons. For example, consider the halfspace or simplicial depth and a random sample of n points from some continuous probability distribution on $\mathbb{R}^d$. Any point lying on the border of convex hull of the n points has empirical halfspace depth equal to $1/n$. Proportion of such points increases with increasing dimension d. Consequently proportion of points with equal empirical depth $(1/n)$ is increasing. Empirical depth of points does not provide much information in such a situation.

To illustrate the problem we performed a simple simulation. 1000 times n points from standardized d-dimensional normal distribution $N_d(\mathbf{0}, \boldsymbol{I})$ were randomly sampled. Number of points lying on the border of the convex hull was counted for each random sample. Subsequently the median of the number of points lying on the border of the convex hull was estimated. Table 1.1 shows estimated medians for $d = 2, 3, 4, 5, 10$ and $n = 20, 50, 100$ and 500. We used the quickhull algorithm implemented in R-library `geometry`. The algorithm did not converged for $d = 10$ and $n = 500$.

dimension (d)	number of sampled points (n)			
	20	50	100	500
2	7	8	9	11
3	12	17	22	32
4	16	28	39	70
5	19	37	57	124
10	20	50	98	

Table 1.1: Median of the number of points lying on the border of the convex hull of the random sample from $N_d(\mathbf{0}, \boldsymbol{I})$.

Chapter 2

Weighted data depth

In Section 1.2.1 of the previous chapter we summarized properties of the halfspace depth, which is one of the most important depth functions. In this chapter an alternative definition of the data depth, which generalises the concept of the halfspace depth, is discussed. This generalization was motivated by some weak points of the halfspace depth, mainly by the convexity of its central regions. This property may be considered as a disadvantage of the halfspace depth (and of other depth functions) when it is applied to considerably non-convex datasets. We propose depth function derived from the halfspace depth function which, in contrary to the halfspace depth, allows the central regions to be more general than convex. The main idea is to use *weights* (weighted probability) in the halfspace rather than the probability of halfspace. Results presented in this chapter were published in [22].

2.1 Definition of the weighted halfspace depth

Let us denote by $\boldsymbol{x}$ the point for which the depth is computed and by $\mathbb{H} \subset \mathbb{R}^d$ the halfspace of interest. Each point $\boldsymbol{y} \in \mathbb{H}$ is assigned a weight $w(\boldsymbol{y})$ which depends on a position of $\boldsymbol{y}$ with respect to $\boldsymbol{x}$ and then the weighted probability $p_{\mathbb{H}} = \int_{\mathbb{H}} w(\boldsymbol{Y})\,\mathrm{dP}$ of the halfspace $\mathbb{H}$ is computed. The same weights are used to the opposite halfspace $\mathbb{R}^d \setminus \mathbb{H}$ and $p_{\mathbb{R}^d \setminus \mathbb{H}}$ is calculated. The ratio of these two values is used for definition of the weighted depth (in contrary to the halfspace depth where the opposite halfspace need not to be considered). The idea of this technique is to compare opposite halfspaces. When the considered distribution is symmetric and the point $\boldsymbol{x}$ is near the centre of the symmetry, then the (weighted) probability of any two opposite halfspaces, separated by a hyperplane containing $\boldsymbol{x}$, is similar. In contrary, when the point $\boldsymbol{x}$ lies far from the centre of symmetry, there is a big difference between the (weighted) probabilities of some opposite halfspaces.

The notion of weight function is crucial in our approach. At this point we want to have the broadest possible class of weight functions that can be used in our approach. Later, we are going to make some restrictions on weight functions to obtain some desirable properties like strong consistency (Section 2.4.1) and depth equal to zero out of the convex hull of the support (Section 2.5).

Another important remark is that any closed halfspace $\mathbb{H}$ determined by a hyperplane containing $\boldsymbol{x}$ is isomorphic to the halfspace $\mathbb{H}_0 = \{\boldsymbol{y} : y_d \geq 0\}$. A closed

halfspace determined by a hyperplane containing $\boldsymbol{x}$ and a normal vector $\boldsymbol{u} \in \mathbb{R}^d : \|u\| = 1$ is described as $\mathbb{H} = \left\{\boldsymbol{y} \in \mathbb{R}^d : \boldsymbol{u}^T(\boldsymbol{y} - \boldsymbol{x}) \geq 0\right\}$. A linear transformation $T : \boldsymbol{y} \mapsto \mathbf{A}(\boldsymbol{y} - \boldsymbol{x})$ maps $\mathbb{H}$ to $\mathbb{H}_0$ for a proper orthogonal $d \times d$ matrix $\mathbf{A}$. Hence we can consider a weight function $w_+ : \mathbb{R}^d \to [0, \infty)$, which can be nonzero only on $\mathbb{H}_0$ and compute the weight in $\boldsymbol{y} \in \mathbb{H}$ as a $w_+(\mathbf{A}(\boldsymbol{y} - \boldsymbol{x}))$. On the opposite halfspace, a "counterweight" function is defined.

Formally, we denote $w_+ : \mathbb{R}^d \to [0, \infty)$ any measurable weight function which is bounded and such that

$$w_+(\boldsymbol{x}) = w_+(x_1, \ldots, x_{d-1}, x_d) = 0 \text{ if } x_d < 0,$$

and denote its "counterweight function" as

$$w_-(\boldsymbol{x}) = w_-(x_1, \ldots, x_{d-1}, x_d) = w_+(x_1, \ldots, x_{d-1}, -x_d).$$

The definition of weighted halfspace depth proceeds directly from the ideas described above.

Definition 2.1 *Let $\boldsymbol{X}$ be a random vector and* P *its probability distribution. The weighted depth of a point $\boldsymbol{x}$ with respect to* P *is defined as*

$$\mathrm{D}(\boldsymbol{x}; P) := \inf_{\mathbf{A} \in \mathbb{O}_d} \frac{\mathrm{E_P}\, w_+\big(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})\big)}{\mathrm{E_P}\, w_-\big(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})\big)}, \tag{2.1}$$

where w_+ is the weight function, $\mathbb{O}_d$ denotes the space of all orthogonal $d \times d$ matrices, and the term $0/0$ is defined to be 1.

Notation remark: In this chapter we will use shorter notation $D(x)$ instead of $D(x; P)$ whenever P is not too important or can be easily identified from the context. The halfspace depth of a point $\boldsymbol{x}$ is denoted by $HD(x)$ in this chapter to distinguish between the halfspace and weighted halfspace depth.

In Definition 2.1 the orthogonal transformations are used to allow full generality of the weight function. For smaller class of symmetric weight functions, in the case when

$$w_+(x_1, \ldots, x_k, \ldots, x_d) = w_+(x_1, \ldots, -x_k, \ldots, x_d), \quad k = 1, \ldots, d-1$$

holds, it is possible to consider only rotations instead of all orthogonal transformations. In particular, the role of the orthogonal transformation is the same as the role of rotations (directions $\boldsymbol{u}$) of the halfspace in Definition 1.3. In other words, instead of rotating the weight function w_+ the random vector $\boldsymbol{X}$ is orthogonally transformed ("rotated to a direction").

Let us examine the possible range of a weighted depth function. Recall that the halfspace depth of any point with respect to a continuous distributions can not be greater than one half and $HD(\boldsymbol{x}) = 1/2$ iff $\boldsymbol{x}$ is the centre of symmetry. To be exact, we should add that the halfspace depth can be greater that $1/2$ when considering discrete distributions. In an extreme case of a distribution concentrated in one point, the halfspace depth of this point is equal to one.

Theorem 2.1 *For any d-dimensional random vector* $\boldsymbol{X}$ *and any* $\boldsymbol{x} \in \mathbb{R}^d$ *it holds* $\mathrm{D}(\boldsymbol{x}) \leq 1$.

Proof: Denote $\mathbf{I}_- = \mathrm{diag}_d(1, 1, \ldots, -1)$ a $d \times d$ diagonal orthogonal matrix. It is not difficult to see that $w_-(\boldsymbol{X}) = w_+(\mathbf{I}_-\boldsymbol{X})$ and $w_+(\boldsymbol{X}) = w_-(\mathbf{I}_-\boldsymbol{X})$. Since $\{\mathbf{I}_-\mathbf{A} : \mathbf{A} \in \mathbb{O}_d\} = \mathbb{O}_d$ it follows

$$\mathrm{D}(\boldsymbol{x}) = \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\mathrm{E}w_+(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x}))}{\mathrm{E}w_-(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x}))} = \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\mathrm{E}w_-(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x}))}{\mathrm{E}w_+(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x}))} \tag{2.2}$$

and since clearly

$$\min\left\{\frac{\mathrm{E}\, w_+(\boldsymbol{Y})}{\mathrm{E}\, w_-(\boldsymbol{Y})}, \frac{\mathrm{E}\, w_-(\boldsymbol{Y})}{\mathrm{E}\, w_+(\boldsymbol{Y})}\right\} \leq 1$$

the proof is completed.

□

The connection between the depth function of Definition 2.1 and the halfspace depth function need not be clear at this moment. In the following discussion it is shown that the depth function D is essentially a generalisation of the halfspace depth.

Definition 2.2 *Define a depth function*

$$\widetilde{\mathrm{D}}(\boldsymbol{x}) := \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\mathrm{E}w_+(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x}))}{\mathrm{E}w_+(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})) + \mathrm{E}w_-(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x}))}, \tag{2.3}$$

for a weight function w_+*; the ratio* $0/(0+0)$ *is now defined as* $1/2$.

The depth functions D and $\widetilde{\mathrm{D}}$ are equivalent in the sense of the multivariate ordering:

Theorem 2.2 *For any weight function* w_+ *and for all* $\boldsymbol{x}$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2 \in \mathbb{R}^d$ *the equivalence*

$$\mathrm{D}(\boldsymbol{x}_1) \leq \mathrm{D}(\boldsymbol{x}_2) \iff \widetilde{\mathrm{D}}(\boldsymbol{x}_1) \leq \widetilde{\mathrm{D}}(\boldsymbol{x}_2) \tag{2.4}$$

holds. Moreover,

$$\widetilde{\mathrm{D}}(\boldsymbol{x}) \leq \frac{1}{2}, \tag{2.5}$$

and

$$\mathrm{D}(\boldsymbol{x}) = \frac{\widetilde{\mathrm{D}}(\boldsymbol{x})}{1-\widetilde{\mathrm{D}}(\boldsymbol{x})}. \tag{2.6}$$

Proof: Following similar argument as in proof of Theorem 2.1 it holds

$$\begin{aligned}\widetilde{\mathrm{D}}(\boldsymbol{x}) &= \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\mathrm{E}_{\mathrm{P}}\, w_+\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big)}{\mathrm{E}_{\mathrm{P}}\, w_+\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big) + \mathrm{E}_{\mathrm{P}}\, w_-\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big)} \\ &= \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\mathrm{E}_{\mathrm{P}}\, w_-\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big)}{\mathrm{E}_{\mathrm{P}}\, w_+\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big) + \mathrm{E}_{\mathrm{P}}\, w_-\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big)}.\end{aligned}$$

The inequality (2.5) follows from the obvious fact that

$$\min\left\{\frac{\mathrm{E}\, w_+(\boldsymbol{Y})}{\mathrm{E}\, w_+(\boldsymbol{Y})+\mathrm{E}\, w_-(\boldsymbol{Y})}, \frac{\mathrm{E}\, w_-(\boldsymbol{Y})}{\mathrm{E}\, w_+(\boldsymbol{Y})+\mathrm{E}\, w_-(\boldsymbol{Y})}\right\} \le 1/2.$$

Denote for fixed orthogonal matrix $\mathbf{A}$

$$v_+ = \mathrm{E} w_+\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big) \text{ and } \; v_- = \mathrm{E} w_-\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big).$$

If $v_- > 0$ then

$$\frac{v_+}{v_-} = \frac{v_+}{v_-+v_+}\left(\frac{v_-}{v_-+v_+}\right)^{-1} = \frac{v_+}{v_-+v_+}\left(1-\frac{v_+}{v_-+v_+}\right)^{-1}. \tag{2.7}$$

If $v_- = 0$ and $v_+ > 0$ then v_- and v_+ in (2.7) may be interchanged (see arguments for (2.5) and (2.2)).

If both $v_- = v_+ = 0$ then the $0/0$ ratios are defined as

$$\frac{v_+}{v_-} = 1, \; \frac{v_+}{v_-+v_+} = \frac{1}{2} \Rightarrow \frac{v_+}{v_-} = \frac{v_+}{v_-+v_+}\left(1-\frac{v_+}{v_-+v_+}\right)^{-1}.$$

Equation(2.6) now follows.

Since the function $x \mapsto x/(1-x)$ is increasing in x for $x \in [0, 1/2]$, the equivalence (2.4) follows.

□

The previous theorem shows that our definition is in some sense a direct generalisation of the halfspace depth if the underlying distribution is absolutely continuous. Indeed, the halfspace depth $\mathrm{HD}(\boldsymbol{x})$ is equal to $\widetilde{\mathrm{D}}(\boldsymbol{x})$ for $w_+(\boldsymbol{y}) \equiv 1$ (the denominator is 1 for any absolutely continuous distribution).

In the case of non-continuous distribution, it holds $\mathrm{HD}(\boldsymbol{x}) \ge \widetilde{\mathrm{D}}(\boldsymbol{x})$ for all $\boldsymbol{x}$ and the inequality may be strict at some points. Indeed, consider $p \in (0,1)$ and a bivariate distribution given by $(1-p)\mathrm{Unif}_{[0,1]^2} + p\delta_{(1,1)}$ - the mixture of the uniform distribution on $[0,1]^2$ and a point mass at $(1,1)$ (see Figure 2.1). Then, obviously, $\mathrm{HD}(1,1) = p > p/(1+p) = \widetilde{\mathrm{D}}(1,1)$.

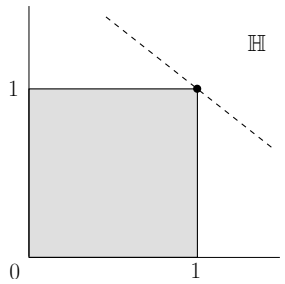


Figure 2.1: $\mathrm{HD}(\boldsymbol{x})$ might be greater than $\widetilde{\mathrm{D}}(\boldsymbol{x})$ for non-continuous distributions.

Obviously, the empirical measure P_n is used for the definition of the sample weighted depth. In this chapter we shall call $\mathrm{D}(\boldsymbol{x})$ simply *the depth of* $\boldsymbol{x}$ unless we need to distinguish more depth functions.

We find as a natural choice of weight function a function which is *spherically symmetric* about x_d-axis. It means that there exists function $h:\ [0,+\infty)\times\mathbb{R}\to\mathbb{R}$ such that

$$w_+(x_1,\ldots,x_d)=h(x_1^2+\ldots+x_{d-1}^2,x_d).$$

Namely, it holds $w_+(\boldsymbol{x})=w_+(x_1,\ldots,x_{d-1},x_d)=w_+(-x_1,\ldots,-x_{d-1},x_d)=w_-(-\boldsymbol{x})$ in this case.

2.2 Examples of weight function

In this section we want to introduce several possibilities how to choose the weight function w_+.

1. The **band weight function** is for a chosen $h>0$ defined as

$$w_+(x_1,\ldots,x_d)=\begin{cases}1 & \text{if } \sum_{i=1}^{d-1}x_i^2<h^2,\ x_d>0,\\ 0 & \text{elsewhere.}\end{cases} \tag{2.8}$$

   In particular, for $\mathbb{R}^2$ the definition has the following meaning. Given fixed point $\boldsymbol{x}$ and a direction $\boldsymbol{s}$ (unit vector in $\mathbb{R}^2$), we consider a line $l=\boldsymbol{x}+t\boldsymbol{s},\ t\in\mathbb{R}$ for which a band with width $2h$

$$\mathbb{B}(\boldsymbol{x},\boldsymbol{s})=\{\boldsymbol{y}\in\mathbb{R}^2:\ d(\boldsymbol{y},l)<h\}$$

   is defined (d denotes the Euclidean distance). The band $\mathbb{B}(\boldsymbol{x},\boldsymbol{s})$ is divided by a segment orthogonal to $\boldsymbol{s}$ and containing $\boldsymbol{x}$ into two half-bands $\mathbb{B}_+(\boldsymbol{x},\boldsymbol{s})$ and $\mathbb{B}_-(\boldsymbol{x},\boldsymbol{s})$. Denoting $p_+(\boldsymbol{x},\boldsymbol{s})$ and $p_-(\boldsymbol{x},\boldsymbol{s})$ the probabilities of $\mathbb{B}_+(\boldsymbol{x},\boldsymbol{s})$ and $\mathbb{B}_-(\boldsymbol{x},\boldsymbol{s})$ respectively, the (*band*) weighted depth becomes

$$\mathrm{D}(\boldsymbol{x})=\inf_{\|\boldsymbol{s}\|=1}\frac{p_+(\boldsymbol{x},\boldsymbol{s})}{p_-(\boldsymbol{x},\boldsymbol{s})}.$$

   The example is illustrated in Figure 2.2 (left). It is clear, that broadening the band by letting h go to infinity leads to the depth which is equivalent to the halfspace depth for continuous distributions. The sample version is calculated from the number of observations in $\mathbb{B}_+(\boldsymbol{x},\boldsymbol{s})$ and $\mathbb{B}_-(\boldsymbol{x},\boldsymbol{s})$.

2. The **cone weight function** is defined for an angle $\alpha\in(0,\pi/2]$ as

$$w_+(x_1,\ldots,x_d)=\begin{cases}1 & \text{if } \angle\big((x_1,\ldots,x_d),(0,\ldots,0,x_d)\big)\le\alpha\\ 0 & \text{elsewhere,}\end{cases}$$

   where $\angle(\boldsymbol{x},\boldsymbol{y})$ denotes the angle between two vectors. The example in two dimensional case is illustrated in Figure 2.2 (right). Clearly, for a continuous distribution and $\alpha=\pi/2$ the depth is equivalent to the halfspace depth.

   In some sense the cone weight function is a *modification* of cylinder weight function. To see that, it is sufficient to use an appropriate function $h(x_d)$ instead of a constant h in definition (2.8).

3. The **local weight function** is for a chosen $r > 0$ defined as

$$w_+(x_1, \ldots, x_d) = \begin{cases} 1 & \text{if } \sum_{i=1}^{d} x_i^2 < r \\ 0 & \text{elsewhere.} \end{cases}$$

This choice seems to be reasonable, emphasizing properties of the distribution near the point of interest $\boldsymbol{x}$. However, this weight function is inadvisable at all. Consider a point $\boldsymbol{x}$ far enough from the support of distribution. Then there is no probability mass in the sphere around the point $\boldsymbol{x}$ and the local weighted depth of the point is $0/0 = 1$.

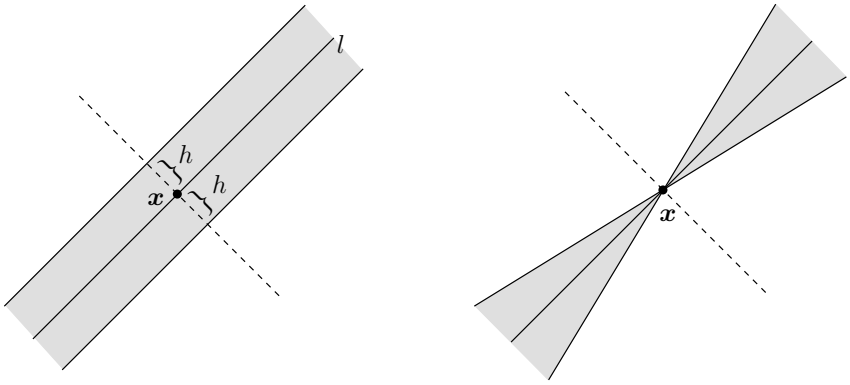


Figure 2.2: Scheme of possible weight functions - band weight function (left), cone weight function (right).

In the first tree examples, we considered weight functions, that are constant on some segments of the halfspace. Nevertheless, the class of possible weight functions is much broader. We introduce some other possibilities.

4. The **ridge weight function** is for a chosen $h > 0$ defined as

$$w_+(x_1, \ldots, x_d) = \begin{cases} 1 - \sqrt{\sum_{i=1}^{d-1} x_i^2}/h & \text{if } \sqrt{\sum_{i=1}^{d-1} x_i^2} < h,\ x_d > 0, \\ 0 & \text{elsewhere.} \end{cases}$$

5. The **normal weight function** is defined as

$$w_+(x_1, \ldots, x_d) = \begin{cases} \phi_{\boldsymbol{\Sigma}}(x_1, \ldots, x_{d-1}) & \text{if } x_d > 0 \\ 0 & \text{elsewhere,} \end{cases}$$

where $\phi_{\boldsymbol{\Sigma}}$ is the density of $d-1$ dimensional normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$. It is, however, also possible to generalise the weight function in the way that the matrix $\boldsymbol{\Sigma}$ may be a function of x_d.

The Figure 2.3 compares the band weight function (example 1) with the ridge weight function (example 4) and the normal weight function (example 5). The value of these weight functions does not depend on the x_d. Figure 2.3 displays the considered weight functions as the functions of the first coordinate x_1.

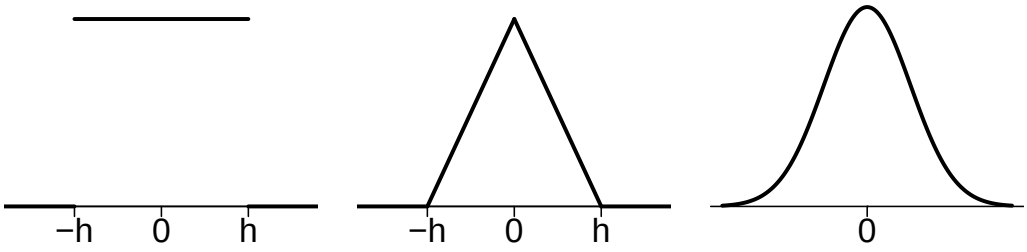


Figure 2.3: Band weight (left), ridge weight (middle) and normal weight (right) as functions of x_1.

2.3 Basic properties of weighted depth

Let us summarize some facts about the depth function D. We focus on the properties of depth function (stated in Definition 1.1). In general the function D does not fulfill all these properties. For example, D is not generally affine invariant. Only translation and rotation invariance can be proved.

Theorem 2.3 *The depth function defined by* (2.1) *is translation invariant.*

Proof: It follows directly from the definition that

$$\mathrm{D}(\boldsymbol{x}+\boldsymbol{a};\mathrm{P}_{\boldsymbol{X}+\boldsymbol{a}}) = \mathrm{D}(\boldsymbol{x};\mathrm{P}_{\boldsymbol{X}}).$$

□

Theorem 2.4 *The depth function defined by* (2.1) *is rotation invariant.*

Proof: Every rotation of a vector $\boldsymbol{x} \in \mathbb{R}^d$ may be written as $\mathbf{B}\boldsymbol{x}$, where $\mathbf{B} \in \mathbb{O}_d$ is some orthogonal $d \times d$ matrix. Hence,

$$\begin{aligned}\mathrm{D}(\mathbf{B}\boldsymbol{x};\mathrm{P}_{\mathbf{B}\boldsymbol{X}}) &= \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\mathrm{E_P}\, w_+\big(\mathbf{A}(\mathbf{B}\boldsymbol{X}-\mathbf{B}\boldsymbol{x})\big)}{\mathrm{E_P}\, w_-\big(\mathbf{A}(\mathbf{B}\boldsymbol{X}-\mathbf{B}\boldsymbol{x})\big)} = \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\mathrm{E_P}\, w_+\big(\mathbf{AB}(\boldsymbol{X}-\boldsymbol{x})\big)}{\mathrm{E_P}\, w_-\big(\mathbf{AB}(\boldsymbol{X}-\boldsymbol{x})\big)} \\ &= \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\mathrm{E_P}\, w_+\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big)}{\mathrm{E_P}\, w_-\big(\mathbf{A}(\boldsymbol{X}-\boldsymbol{x})\big)} = \mathrm{D}(\boldsymbol{x};\mathrm{P}_{\boldsymbol{X}})\end{aligned}$$

since $\{\mathbf{AB}:\ \mathbf{A}\in\mathbb{O}_d\} = \mathbb{O}_d$ as follows from the orthogonality of $\mathbf{B}$.

□

The depth need not decrease along a ray from the deepest point (even if the deepest point is unique). And the sets

$$\{\boldsymbol{x} : \mathrm{D}(\boldsymbol{x}) \geq d\}, \quad d \in [0,1] \tag{2.9}$$

need not be convex and may be sometimes disconnected. This fact depends on the underlying distribution; however, in some situations these properties are desirable. There is a particular interest in the so called *deepest point*, i.e., the point $\widetilde{\boldsymbol{x}}$ for which

$$\mathrm{D}(\widetilde{\boldsymbol{x}}) = \max_{\boldsymbol{x}} \mathrm{D}(\boldsymbol{x}).$$

Definition (2.1) in general does not give a unique deepest point even in a situation of an absolutely continuous distribution with connected support.

Let us consider the uniform distribution on a set

$$S = \{(x_1, x_2)^T : \ 0 \le x_1 \le 10, \ 0 \le x_2 \le 1\} \cup \{(x_1, x_2) : \ 0 \le x_1 \le 1, \ 0 \le x_2 \le 10\}.$$

Let us consider the band weight function (2.8) with a small h, say $h = 1/20$, and the corresponding weighted depth function. From the shape of the support S it follows that the only unique deepest point may lie on a line $x_1 = x_2$ only. It can be seen that for any point $\boldsymbol{x}$ on the line $x_1 = x_2$ it holds $\mathrm{D}(\boldsymbol{x}) \le 1/9$.

Consider the point $\boldsymbol{z} = (5, 1/2)$. After some calculations we get $\mathrm{D}(\boldsymbol{z}) > 1/9 \ge \mathrm{D}(\boldsymbol{x})$ for any $\boldsymbol{x} = (x_1, x_1)^T$. Indeed, the lower estimate for $\mathrm{D}(\boldsymbol{z})$ may be obtained considering a line l connecting $\boldsymbol{z}$ and the point $(0, 10)$ together with a band b of the width $2h$ around l and, on the other hand considering a line l' connecting $\boldsymbol{z}$ and the point $(5, 0)$ with the same band around. See Figure 2.4 for a visualisation of this example.

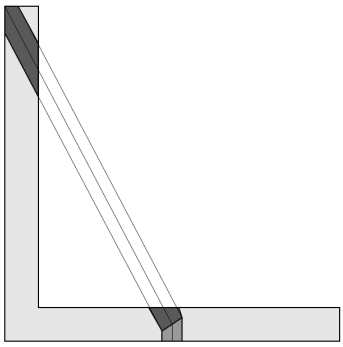


Figure 2.4: The deepest point need not to be unique.

In this example there is no natural central point although the distribution is symmetric about the line $x_2 = x_1$. There are two deepest points (symmetric about the line of symmetry). The central regions are symmetric about the $x_1 = x_2$ axis as well.

In the previous example there is not any "natural" deepest point. On the other hand, if there is an intuitive deepest point, like the point of central symmetry, we would like to prove that it is the deepest point for the weighted depth function. Indeed it is the case for a suitable weight function. In what follows we use notions of symmetry as they were recalled in Section 1.1.

Theorem 2.5 *Let w_+ be such that $w_+(x_1, \ldots, x_{d-1}, x_d) = w_+(-x_1, \ldots, -x_{d-1}, x_d)$ (is symmetric about x_d-axis) and suppose that the distribution of $\boldsymbol{X}$ is centrally symmetric about point $\boldsymbol{\theta}$. Then*

$$\mathrm{D}(\boldsymbol{x}) \le \mathrm{D}(\boldsymbol{\theta}) = 1, \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Proof: It can be assumed that $\boldsymbol{\theta} = \mathbf{0}$ without loss of generality (translation invariance of D). Since w_+ is symmetric about x_d-axis and $w_-(\boldsymbol{x}) = w_+(\mathbf{I}_-\boldsymbol{x})$ it holds

$$w_+(\boldsymbol{x}) = w_-(-\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

It follows that

$$\mathrm{E}\, w_-(\mathbf{A}\boldsymbol{X}) = \mathrm{E}\, w_+(-\mathbf{A}\boldsymbol{X}) = \mathrm{E}\, w_+(\mathbf{A}\boldsymbol{X})$$

for $\boldsymbol{X}$ centrally symmetric about $\mathbf{0}$ and arbitrary matrix $\mathbf{A} \in \mathbb{O}_d$. Thus $\mathrm{D}(\mathbf{0}) = 1$. The fact that $\mathrm{D}(\boldsymbol{x}) \leq 1,\ \forall \boldsymbol{x}$ completes the proof.

□

This result may be extended to angular symmetric distributions.

Theorem 2.6 *Let w_+ be symmetric about x_d-axis and suppose that the distribution of $\boldsymbol{X}$ is angular symmetric about point $\boldsymbol{\theta}$. If w_+ is such that*

$$w_+(k\boldsymbol{x}) = w_+(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathbb{R}^d,\ k \geq 0 \tag{2.10}$$

then

$$\mathrm{D}(\boldsymbol{x}) \leq \mathrm{D}(\boldsymbol{\theta}) = 1, \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Proof: It is an analogue to the proof of Theorem 2.5. Let $\boldsymbol{\theta} = \mathbf{0}$ without loss of generality. Under the assumption (2.10) it holds

$$\begin{aligned}\mathrm{E}\, w_-(\mathbf{A}\boldsymbol{X}) &= \mathrm{E}\, w_+(-\mathbf{A}\boldsymbol{X}) = \mathrm{E}\, w_+(-\mathbf{A}\boldsymbol{X}/\|\mathbf{A}\boldsymbol{X}\|) \\ &= \mathrm{E}\, w_+(\mathbf{A}\boldsymbol{X}/\|\mathbf{A}\boldsymbol{X}\|) = \mathrm{E}\, w_+(\mathbf{A}\boldsymbol{X})\end{aligned}$$

$\forall \mathbf{A} \in \mathbb{O}_d$, hence

$$\mathrm{D}(\mathbf{0}) = \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\mathrm{E}\, w_+(\mathbf{A}\boldsymbol{X})}{\mathrm{E}\, w_-(\mathbf{A}\boldsymbol{X})} = 1.$$

□

In Theorem 2.6 it is sufficient to define the weight function w_+ on the unit halfsphere

$$S_{d,+} = \{\boldsymbol{x} :\ \|\boldsymbol{x}\| = 1,\ x_d \geq 0\}$$

and use $w_+(\boldsymbol{x}) = w_+(\boldsymbol{x}/\|\boldsymbol{x}\|)$ to ensure (2.10). Obviously the cylinder (band) depth does not satisfy the assumption of Theorem 2.6. On the other hand the assumption of the theorem is satisfied by the cone weight function defined in example 2 of Section 2.2.

2.4 Consistency of the depth function

We shall prove in this section a strong pointwise consistency of the depth function under relatively mild conditions on the weight function. Note that the consistency of the halfspace depth is a direct corollary to our result.

In what follows we consider an *absolutely continuous* Borel probability measure P on $\mathbb{R}^d$. Let us denote

$$\angle(\boldsymbol{u}, \boldsymbol{v}) \text{ the angle of vectors } \boldsymbol{u} \text{ and } \boldsymbol{v},$$

and

$$\mathbb{A}_\varphi \subset \mathbb{O}_d \text{ set of all rotation matrices } \mathbf{A} \text{ such that } \angle(\boldsymbol{u}, \mathbf{A}\boldsymbol{u}) \leq \varphi \text{ for all } \boldsymbol{u} \in \mathbb{R}^d.$$

Note that $\mathbb{A}_0 = \{\mathbf{I}_d\}$. Finally, let us denote by $\mathbf{N}_{\boldsymbol{s}}$ any matrix representing an *orthogonal rotation* such that $\mathbf{N}_{\boldsymbol{s}}\boldsymbol{s} = (\mathbf{0}^T, 1)^T$, $\mathbf{N}_{(\mathbf{0}^T,1)^T} := \mathbf{I}_d$. Such a matrix need not to be defined uniquely, however, for any two different $\mathbf{N}^1_{\boldsymbol{s}},\ \mathbf{N}^2_{\boldsymbol{s}}$ it holds

$$\angle(\mathbf{N}^1_{\boldsymbol{s}}\boldsymbol{u}, \mathbf{N}^1_{\boldsymbol{s}}\boldsymbol{v}) = \angle(\mathbf{N}^2_{\boldsymbol{s}}\boldsymbol{u}, \mathbf{N}^2_{\boldsymbol{s}}\boldsymbol{v}) = \angle(\boldsymbol{u}, \boldsymbol{v}) \quad \text{for all} \quad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d.$$

2.4.1 Regularity of weight function

The class of weight functions w_+ used in the definition 2.1 is too broad. To be able to prove a strong pointwise consistency of the depth function, at least some "regularity" properties of the weight function are needed. We tried to find properties that enable to prove a strong consistency, but that are not too restrictive.

Definition 2.3 *We say that weight function w_+ satisfies regularity conditions if*

(A) $w_+(x_1, \ldots, x_{d-1}, x_d)$ *is spherically symmetric about x_d-axis, i.e. w_+ is a function of $(x_1^2 + \cdots + x_{d-1}^2, x_d)$. In other words, w_+ is a function of the distance from x_d axis and values on x_d axis.*

(B) w_+ *is measurable and bounded.*

(C) For arbitrary point $\boldsymbol{x}$ it holds that

$$\lim_{\varphi \to 0+} \sup_{\mathbf{A} \in \mathbb{A}_\varphi} \Big\{ w_+\big(\mathbf{A}\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{X} - \boldsymbol{x})\big) \Big\} = w_+\big(\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{X} - \boldsymbol{x})\big) \text{ P-}a.s.$$

$$\lim_{\varphi \to 0+} \inf_{\mathbf{A} \in \mathbb{A}_\varphi} \Big\{ w_+\big(\mathbf{A}\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{X} - \boldsymbol{x})\big) \Big\} = w_+\big(\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{X} - \boldsymbol{x})\big) \text{ P-}a.s.$$

for every direction $\boldsymbol{s}$, $\|\boldsymbol{s}\| = 1$. In other words, the sup*, resp.* inf *function over all orthogonal rotations is* P*-a.s. continuous from right in* 0 *with respect to a rotation angle.*

Let us first denote two important subsets of points. Define

$$\begin{aligned}
\mathcal{H}_1 =& \{\boldsymbol{x} : \inf_{\mathbf{A} \in \mathbb{O}_d} \mathrm{E}\, w_+(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})) > 0\}, \\
\mathcal{H}_2 =& \{\boldsymbol{x} : \exists \delta > 0\ \forall \varepsilon > 0\ \exists \mathbf{A}_\varepsilon \in \mathbb{O}_d : \mathrm{E}\, w_+(\mathbf{A}_\varepsilon(\boldsymbol{X} - \boldsymbol{x})) < \varepsilon \text{ and} \\
& \mathrm{E}\, w_-(\mathbf{A}_\varepsilon(\boldsymbol{X} - \boldsymbol{x})) > \delta\}.
\end{aligned}$$

Recall now the notion of probability support and closed convex support. The *support* $\mathrm{sp(P)}$ of probability measure P is the smallest closed set with probability 1, i.e.

$$\mathrm{sp(P)} = \bigcap \{F \in \mathcal{F} : \mathrm{P}(F) = 1\},$$

where $\mathcal{F}$ denotes class of all closed subsets. The *closed convex support* $\mathrm{csp(P)}$ of probability measure P is defined as closed convex hull of the support $\mathrm{sp(P)}$.

It is easy to see that the set $\mathcal{H}_1$ contains the interior of support $\mathrm{sp(P)}$, i.e. points whose open neighbourhood is contained in the support of P. In the case of absolutely continuous distribution $\mathrm{P}(\mathcal{H}_1) = 1$. On the other hand the set $\mathcal{H}_2$ represents points with zero depth and, in particular, for the *complement* of closed convex support $\complement\mathrm{csp(P)}$ it holds $\complement\mathrm{csp(P)} \subset \mathcal{H}_2$ under very weak conditions on the weight function w_+. It is easy to see that if $\boldsymbol{x} \in \mathcal{H}_1$ then $\mathrm{D}(\boldsymbol{x}; \mathrm{P}) > 0$ and if $\boldsymbol{x} \in \mathcal{H}_2$ then $\mathrm{D}(\boldsymbol{x}; \mathrm{P}) = 0$.

2.4.2 Empirical process theory - recall

Before we state the main claim of this section, we want to recall some notions from empirical process theory. The lemma is stated without proof, which can be found in [53].

First, define a metric on some class of functions $\mathcal{G}$.

Definition 2.4 ***$L_p(Q)$-metric***
Let Q be a measure on $(\mathcal{X}, \mathcal{A})$ and p a real constant such that $1 \leq p < \infty$. Denote by $L_p(Q)$ the set of all real functions whose p-th power is absolutely integrable with respect to the measure Q:

$$L_p(Q) = \left\{ g : \mathcal{X} \to \mathbb{R} : \int |g|^p \, \mathrm{dQ} < \infty \right\}.$$

For $g \in L_p(Q)$ define

$$\|g\|_{p,Q}^p = \int |g|^p \, \mathrm{dQ}.$$

We refer to $\|\cdot\|_{p,Q}$ as the $L_p(Q)$-metric.

In our case we will consider the probability measure P and the $L_1(P)$-metric.

Definition 2.5 ***Entropy with bracketing for the $L_p(Q)$-metric***
Let $N_{p,B}(\delta, \mathcal{G}, Q)$ be the smallest value of N for which there exist pairs of functions $\left\{ \left(g_j^L, g_j^U\right) \right\}_{j=1}^N$ such that $\left\| g_j^U - g_j^L \right\|_{p,Q} \leq \delta$ for all $j = 1, \ldots, N$, and such that for each $g \in \mathcal{G}$, there exist $j \in \{1, \ldots, N\}$ such that

$$g_j^L \leq g \leq g_j^U.$$

Then $H_{p,B}(\delta, \mathcal{G}, Q) = \log N_{p,B}(\delta, \mathcal{G}, Q)$ is called the δ-entropy with bracketing of $\mathcal{G}$.

Definition 2.6 ***Uniform Law of Large Numbers***
We say that the class of functions $\mathcal{G}$ satisfies the Uniform Law of Large Numbers if

$$\sup_{g \in \mathcal{G}} \left| \int g \; \mathrm{d(P_n - P)} \right| \to 0 \quad \text{-a.s.}$$

Lemma 2.7 *Suppose that*

$$H_{p,B}(\delta, \mathcal{G}, Q) < \infty \quad \text{for all} \delta > 0.$$

Then $\mathcal{G}$ satisfies the Uniform Law of Large Numbers.

2.4.3 Strong pointwise consistency of the depth function

Theorem 2.8 *Let P_n be an empirical measure defined by a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from distribution P. Let the weight function w_+ satisfies the regularity conditions of Definition 2.3. Then for any $\boldsymbol{x} \in \mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ it holds*

$$\mathrm{D}(\boldsymbol{x}; \mathrm{P}_n) \to \mathrm{D}(\boldsymbol{x}; \mathrm{P}) \quad \mathrm{P}\text{-almost surely.} \tag{2.11}$$

Proof: For our purposes we will use standard conventions from measure theory for the extended real line $[-\infty, +\infty]$, e.g. $0.(\pm\infty) = 0$, $+\infty + \infty = +\infty$, etc. and we define logarithm in zero: $\log 0 := \lim_{x\to 0+} \log x = -\infty$.

The first step is to show that the class of functions $\mathcal{W} := \{\boldsymbol{y} \mapsto w_+(\mathbf{A}(\boldsymbol{y}-\boldsymbol{x})) : \mathbf{A} \in \mathbb{O}_d\}$ satisfies the *Uniform law of large numbers*. It means to prove that

$$\sup_{\mathbf{A}\in\mathbb{O}_d} \Big| \frac{1}{n} \sum_{i=1}^{n} w_+(\mathbf{A}(\boldsymbol{X}_i - \boldsymbol{x})) - \mathrm{E_P}\, w_+(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})) \Big| \;\longrightarrow\; 0 \quad \text{P-a.s.} \tag{2.12}$$

To this end it is sufficient to prove that

$$H_{1,B}(\varepsilon, \mathcal{W}, \mathrm{P}) < +\infty, \quad \text{for all } \varepsilon > 0,$$

where $H_{1,B}(\varepsilon, \mathcal{W}, \mathrm{P})$ denotes entropy with ε-bracketing for $L_1(\mathrm{P})$-metric (see Lemma 2.7).

For a fixed vector $\boldsymbol{s}$ and a given angle φ we define functions

$$\begin{aligned} W^U_{\boldsymbol{s},\varphi}(\boldsymbol{z}) &= \sup\{w_+(\mathbf{A}\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{z}-\boldsymbol{x})) :\ \mathbf{A} \in \mathbb{A}_\varphi\}, \\ W^L_{\boldsymbol{s},\varphi}(\boldsymbol{z}) &= \inf\{w_+(\mathbf{A}\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{z}-\boldsymbol{x})) :\ \mathbf{A} \in \mathbb{A}_\varphi\}. \end{aligned}$$

Since $\mathbb{A}_0 = \{\mathbf{I}_p\}$ it holds $W^L_{\boldsymbol{s},0}(\boldsymbol{z}) = W^U_{\boldsymbol{s},0}(\boldsymbol{z}) = w_+(\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{z}-\boldsymbol{x}))$. Further the inequality

$$W^L_{\boldsymbol{s},\varphi}(\boldsymbol{z}) \le w_+(\mathbf{N}_{\boldsymbol{a}}(\boldsymbol{z}-\boldsymbol{x})) \le W^U_{\boldsymbol{s},\varphi}(\boldsymbol{z}) \tag{2.13}$$

holds for arbitrary $\boldsymbol{z}$ and direction $\boldsymbol{a}$ such that $\angle(\boldsymbol{a}, \boldsymbol{s}) \le \varphi$.

For arbitrary direction $\boldsymbol{s}$ we define function

$$G_{\boldsymbol{s}}(\varphi) = \mathrm{E_P} W^U_{\boldsymbol{s},\varphi}(\boldsymbol{X}).$$

This definition is correct, because for a measurable function w_+, the function $W^U_{\boldsymbol{s},\varphi}(\boldsymbol{z})$ is (universally) measurable; see Lemma 2.9 and its proof.

We will show that $G_{\boldsymbol{s}}$ is continuous from right in 0. Since w_+ is bounded, one has that $G_{\boldsymbol{s}}(\varphi) < +\infty$ for all $\varphi \in [0, \pi]$. Measurability and integrability together with condition (C) directly imply continuity from right of $G_{\boldsymbol{s}}$ in 0 using Lebesgue's dominated convergence theorem.

It follows that for all $\varepsilon > 0$ there exists φ_0 such that for all $\varphi \in [0, \varphi_0)$ holds

$$\begin{aligned} \varepsilon > |G_{\boldsymbol{s}}(\varphi) - G_{\boldsymbol{s}}(0)| &= \big|\mathrm{E_P}\big[W^U_{\boldsymbol{s},\varphi}(\boldsymbol{X}) - w_+(\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{X}-\boldsymbol{x}))\big]\big| \\ &= \mathrm{E_P}\big|W^U_{\boldsymbol{s},\varphi}(\boldsymbol{X}) - w_+(\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{X}-\boldsymbol{x}))\big|. \end{aligned}$$

Since inequality (2.13) holds, the last equation is correct. An analogous inequality holds for $W^L_{\boldsymbol{s},\varphi}$.

Hence, for arbitrary $\boldsymbol{s}$, $\|\boldsymbol{s}\| = 1$, and for every $\varepsilon > 0$ there exists $\varphi_{\boldsymbol{s}} > 0$ such that

$$\mathrm{E_P}\big|W^U_{\boldsymbol{s},\varphi_{\boldsymbol{s}}}(\boldsymbol{X}) - W^L_{\boldsymbol{s},\varphi_{\boldsymbol{s}}}(\boldsymbol{X})\big| < \varepsilon. \tag{2.14}$$

Now, for arbitrary $\varepsilon > 0$, we construct ε-bracketing for $\mathcal{W}$. Let's consider the metric space $(\mathbb{S}_p, \rho)$, where $\mathbb{S}_p = \{\boldsymbol{s} :\ \|\boldsymbol{s}\| = 1\}$ and ρ is the Euclidean distance metric. Space $(\mathbb{S}_p, \rho)$ is closed and bounded, hence it is compact. For arbitrary $\boldsymbol{s} \in \mathcal{S}$ an angle $\varphi_{\boldsymbol{s}}$ which satisfies (2.14) may be found. Denote by $\mathcal{C}(\boldsymbol{s}, \varphi_{\boldsymbol{s}})$ a set of all $\boldsymbol{u} \in \mathbb{S}_p$ such that $\angle(\boldsymbol{u}, \boldsymbol{s}) < \varphi_{\boldsymbol{s}}$. $\mathcal{C}(\boldsymbol{s}, \varphi_{\boldsymbol{s}})$ are open sets in the metric space

$(\mathbb{S}_p, \rho)$ and form an open cover of $\mathbb{S}_p$. Since $\mathbb{S}_p$ is compact it follows that for any open cover there exists a finite subcover. In other words there exists a finite subset U of $\mathbb{S}_p$ such that

$$\mathbb{S}_p = \bigcup_{\boldsymbol{u} \in U} \mathcal{C}(\boldsymbol{u}, \varphi_{\boldsymbol{u}}).$$

Every function from $\mathcal{W}$ is determined by a direction $\boldsymbol{s} \in \mathbb{S}_p$ in the sense that for an arbitrary function $v \in \mathcal{W}$ there exists $\boldsymbol{s} \in \mathbb{S}_p$ such that $v(\boldsymbol{y}) = w_+(\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{y} - \boldsymbol{x}))$ and obviously there exists $\boldsymbol{u} \in U$ such that $\boldsymbol{s} \in \mathcal{C}(\boldsymbol{u}, \varphi_{\boldsymbol{u}})$. Hence $W^U_{\boldsymbol{u},\varphi_{\boldsymbol{u}}}$ and $W^L_{\boldsymbol{u},\varphi_{\boldsymbol{u}}}$ are the corresponding bracketing functions which satisfy (2.13) and (2.14).

Finally, we obtain

$$H_{1,B}(\varepsilon, \mathcal{W}, \mathrm{P}) \leq \mathrm{card}(U) < +\infty.$$

and thus (2.12) holds.

Now we can come up to the proof of consistency of depth $\mathrm{D}(\boldsymbol{x}; \mathrm{P}_n)$. It is a consequence of (2.12). Let us use the notation

$$\widehat{\mathrm{D}}_{\mathbf{A}}(\boldsymbol{x}, \mathrm{P}) = \frac{\mathrm{E_P}\, w_+(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x}))}{\mathrm{E_P}\, w_-(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x}))},$$

where the term $0/0$ is defined again as 1.

First the case $\boldsymbol{x} \in \mathcal{H}_1$ is treated. It holds

$$0 < \mathrm{D}(\boldsymbol{x}; \mathrm{P}) \leq \widehat{\mathrm{D}}_{\mathbf{A}}(\boldsymbol{x}, \mathrm{P}) \leq 1/\mathrm{D}(\boldsymbol{x}; \mathrm{P}) < +\infty, \quad \forall \mathbf{A} \in \mathbb{O}_d.$$

It follows from Lemma 2.10 below that

$$\begin{aligned}
|\log \mathrm{D}(\boldsymbol{x}; \mathrm{P}_n) - \log \mathrm{D}(\boldsymbol{x}; \mathrm{P})| &= |\inf_{\mathbf{A} \in \mathbb{O}_d} \log \widehat{\mathrm{D}}_{\mathbf{A}}(\boldsymbol{x}, \mathrm{P}_n) - \inf_{\mathbf{A} \in \mathbb{O}_d} \log \widehat{\mathrm{D}}_{\mathbf{A}}(\boldsymbol{x}, \mathrm{P})| \\
&\leq \sup_{\mathbf{A} \in \mathbb{O}_d} |\log \widehat{\mathrm{D}}_{\mathrm{P}_n}(\boldsymbol{x}, \mathbf{A}) - \log \widehat{\mathrm{D}}_{\mathrm{P}}(\boldsymbol{x}, \mathbf{A})| \\
&\leq \sup_{\mathbf{A} \in \mathbb{O}_d} \Bigg(\Big| \log \frac{1}{n} \sum_{i=1}^n w_+(\mathbf{A}(\boldsymbol{X}_i - \boldsymbol{x})) - \log \mathrm{E_P}\, w_+(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})) \Big| \\
&\qquad + \Big| \log \frac{1}{n} \sum_{i=1}^n w_-(\mathbf{A}(\boldsymbol{X}_i - \boldsymbol{x})) - \log \mathrm{E_P}\, w_-(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})) \Big| \Bigg) \\
&\leq 2 \sup_{\mathbf{A} \in \mathbb{O}_d} \Big| \log \frac{1}{n} \sum_{i=1}^n w_+(\mathbf{A}(\boldsymbol{X}_i - \boldsymbol{x})) - \log \mathrm{E_P}\, w_+(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})) \Big|
\end{aligned} \tag{2.15}$$

almost surely.

Since (2.12) holds it follows that also

$$\sup_{\mathbf{A} \in \mathbb{O}_d} \Big| \log \frac{1}{n} \sum_{i=1}^n w_+(\mathbf{A}(\boldsymbol{X}_i - \boldsymbol{x})) - \log \mathrm{E_P}\, w_+(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})) \Big| \longrightarrow 0 \quad \mathrm{P\text{-a.s.}}$$

From (2.15) one has that

$$|\log \mathrm{D}(\boldsymbol{x}; \mathrm{P}_n) - \log \mathrm{D}(\boldsymbol{x}; \mathrm{P})| \longrightarrow 0 \quad \mathrm{P\text{-a.s.}}$$

So eventually,

$$|\mathrm{D}(\boldsymbol{x};\mathrm{P}_n) - \mathrm{D}(\boldsymbol{x};\mathrm{P})| \longrightarrow 0 \quad \mathrm{P}\text{-a.s.}$$

We shall now consider the case $\mathcal{H}_2$. For $\boldsymbol{x} \in \mathcal{H}_2$ there exists $\delta > 0$ such that for any $\varepsilon > 0$ there exists $\mathbf{A}_\varepsilon$ and for any $\eta > 0$ there exists n_η such that for $n \geq n_\eta$

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} w_+(\mathbf{A}_\varepsilon(\boldsymbol{X}_i - \boldsymbol{x})) < \mathrm{E_P}\, w_+(\mathbf{A}_\varepsilon(\boldsymbol{X}_i - \boldsymbol{x})) + \eta < \varepsilon + \eta, \\ &\frac{1}{n}\sum_{i=1}^{n} w_-(\mathbf{A}_\varepsilon(\boldsymbol{X}_i - \boldsymbol{x})) > \mathrm{E_P}\, w_-(\mathbf{A}_\varepsilon(\boldsymbol{X}_i - \boldsymbol{x})) - \eta > \delta - \eta, \end{aligned} \tag{2.16}$$

holds P=a.s. (see the definition of $\mathcal{H}_2$ and (2.12)). It follows that for $n \geq n_\eta$

$$\begin{aligned} |D(\boldsymbol{x},\mathrm{P}_n) - D(\boldsymbol{x};\mathrm{P})| &= \left| \inf_{\mathbf{A}\in\mathbb{O}_d} \frac{\frac{1}{n}\sum_{i=1}^{n} w_+(\mathbf{A}(\boldsymbol{X}_i - \boldsymbol{x}))}{\frac{1}{n}\sum_{i=1}^{n} w_-(\mathbf{A}(\boldsymbol{X}_i - \boldsymbol{x}))} - 0 \right| \\ &\leq \frac{\frac{1}{n}\sum_{i=1}^{n} w_+(\mathbf{A}_\varepsilon(\boldsymbol{X}_i - \boldsymbol{x}))}{\frac{1}{n}\sum_{i=1}^{n} w_-(\mathbf{A}_\varepsilon(\boldsymbol{X}_i - \boldsymbol{x}))} \\ &< \frac{\varepsilon + \eta}{\delta - \eta}, \end{aligned}$$

and since ε and η may be chosen arbitrary small the proof is completed.

□

The following two technical lemmas are necessary for the proof of consistency.

Lemma 2.9 *Let the weight function w_+ satisfy regularity conditions and consider fixed $\boldsymbol{s}, \|\boldsymbol{s}\| = 1$ and $\varphi \in [0,\pi]$. Then the function*

$$\boldsymbol{z} \mapsto \sup\{w_+(\mathbf{A}\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{z} - \boldsymbol{x})) : \ \mathbf{A} \in \mathbb{A}_\varphi\}$$

is universally measurable.

Proof: The function w_+ may be considered as a function of a distance ($d = \|\boldsymbol{x}\|$) and the "direction" $\boldsymbol{s} = \boldsymbol{x}/\|\boldsymbol{x}\|$ where $\boldsymbol{s} \in \mathbb{S}_d$ is the unit sphere. We use the metric $\rho(\boldsymbol{s},\boldsymbol{z}) = \angle(\boldsymbol{s},\boldsymbol{z})$ for $\boldsymbol{s},\boldsymbol{z} \in \mathbb{S}_d$.

The problem is therefore equivalent to a problem of measurability of a function

$$g(q,s) = \sup\{f(q,z) : \tau(z,s) \leq e\}$$

if $f : [0,+\infty) \times \mathbb{M} \to [0,+\infty)$ is a measurable function, where $(\mathbb{M},\tau)$ is a separable metric space. Denote $B^a = \{(q,z) : \ f(q,z) > a\}$ and note that B^a is a Borel set for any a due to the measurability of f. Denote $C^a := \{(q,s) : \ g(q,s) > a\}$. It is clear that for any q

$$C^a_q = \mathcal{U}_e(B^a_q),$$

where $M_q = \{s : \ (q,s) \in M\}$ denotes the q-section of a set M and $\mathcal{U}_e(N)$ denotes the e-neighbourhood of a set $N \subset \mathbb{M}$. The set C^a is therefore a projection of a Borel set

$$D^{a,e} = \{(q,s,z) \in [0,+\infty) \times \mathbb{M} \times \mathbb{M} : \ (q,z) \in B^a, \ \tau(s,z) \leq e\}$$

into the first two coordinates.

Since the projection of a Borel set is an analytic and hence a universally measurable set it follows that $g(y,x)$ is universally measurable function.

□

If a function g is universally measurable then for any finite Borel measure μ on $[0,+\infty)\times\mathbb{R}$ (in particular for any probability measure) there exist a pair of Borel functions $g_1,\ g_2$ such that $g_1(y,x)\le g(y,x)\le g_2(y,x)$ and $g_2=g_1$ μ-almost surely. Hence the Lebesgue integral of universally measurable function is well defined.

Lemma 2.10 *Consider two bounded functions $f,\ g:\ M\to\mathbb{R}$. Then*

$$\sup\{|f(x)-g(x)|:\ x\in M\}\ge|\inf\{f(x):\ x\in M\}-\inf\{g(x):\ x\in M\}|.$$

Proof: If $\inf f=\inf g$ then it follows immediately because $\sup|f-g|\ge 0$.

If $\inf f>\inf g$ then there exists $\varepsilon_0>0$ such that for all $\varepsilon,\ 0<\varepsilon<\varepsilon_0$, exists $x_g\in M$ which satisfies

$$\inf g\le g(x_g)<\inf g+\varepsilon<\inf f\le f(x_g).$$

Therefor

$$\sup|f-g|\ge|f(x_g)-g(x_g)|\ge|\inf f-g(x_g)|>|\inf f-\inf g|-\varepsilon$$

for all $\varepsilon,\ 0<\varepsilon<\varepsilon_0$ and the proof of Lemma is completed.

□

2.4.4 Discussion of regularity conditions

This section provides discussion of the regularity conditions introduced in the Section 2.4.1. It is quite clear that the most restrictive regularity condition is (C). In the next theorem a simple sufficient condition for (C) is stated. Further we discuss some counterexamples showing that the sample depth need not to be consistent and we propose conditions on the support of the probability measure P and on the weight function w_+ guaranteeing the consistency on $\mathbb{R}^d$.

Theorem 2.11 *Let us have $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ a d-dimensional sample from absolutely continuous probability distribution* P *and suppose spherically symmetric weight function w_+ about x_d axis. Further assume that w_+ is continuous on some connected set $\mathcal{M}\subseteq\mathbb{R}^{d-1}\times[0,+\infty)$ of positive Lebesgue measure and that w_+ is equal to zero on $\mathbb{R}^d\setminus\mathcal{M}$. Then for any $\boldsymbol{x}\in\mathcal{H}=\mathcal{H}_1\cup\mathcal{H}_2$ it holds*

$$\mathrm{D}(\boldsymbol{x};\mathrm{P}_n)\ \longrightarrow\ \mathrm{D}(\boldsymbol{x};\mathrm{P})\quad \mathrm{P}\text{-}a.s.$$

Proof: We need to check the validity of regularity conditions.

Condition (C) for supremum can be equivalently expressed in the form:

$$\lim_{\varphi\to0+}\sup_{\mathbf{A}\in\mathbb{A}_\varphi}\Big\{w_+\big(\mathbf{A}\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{y}-\boldsymbol{x})\big)f(\boldsymbol{y})\Big\}=w_+\big(\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{y}-\boldsymbol{x})\big)f(\boldsymbol{y})$$

for almost all $\boldsymbol{y}$ and for every direction $\boldsymbol{s}$, $\|\boldsymbol{s}\|=1$. f denotes density of probability distribution P. In the following we will use this form of condition (C) and for fixed $\boldsymbol{s}$ we will work with shifted and rotated random vector $\mathbf{N}_{\boldsymbol{s}}(\boldsymbol{X}-\boldsymbol{x})$ instead of random vector $\boldsymbol{X}$. Its density we denote by $f_{\boldsymbol{s}}$.

If $\boldsymbol{y} \notin \mathrm{clo}(\mathcal{M})$ then $\mathbb{R}^d \setminus \mathrm{clo}(\mathcal{M})$ is open set and thus there exists $\varphi_0 > 0$ such that for all $0 \le \varphi < \varphi_0$ it holds that $w_+(\mathbf{B}\boldsymbol{y}) f_{\boldsymbol{s}}(\boldsymbol{y}) = 0$, where $\mathbf{B} \in \mathbb{O}_d$ is arbitrary orthogonal rotation about angle φ.

If $\boldsymbol{y} \in \mathrm{int}(\mathcal{M})$ then, since $\mathrm{int}(\mathcal{M})$ is open and w_+ is there continuous, one has that for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\mathcal{B}(\boldsymbol{y}, \delta) = \{\boldsymbol{u} : \ \|\boldsymbol{u} - \boldsymbol{y}\| < \delta\} \subseteq \mathrm{int}(\mathcal{M})$ and inequality $|w_+(\boldsymbol{u}) - w_+(\boldsymbol{y})| < \varepsilon$ holds for every $\boldsymbol{u} \in \mathcal{B}(\boldsymbol{y}, \delta)$. For every such δ there exists angle $\varphi_0 > 0$ such that for arbitrary rotation $\mathbf{B} \in \mathbb{O}_d$ about angle smaller than φ_0 one has $\mathbf{B}\boldsymbol{y} \in \mathcal{B}(\boldsymbol{y}, \delta)$ and thus $|w_+(\mathbf{B}\boldsymbol{y}) - w_+(\boldsymbol{y})| < \varepsilon$. For any angle $\xi, \ 0 \le \xi < \varphi_0$, we define set

$$\mathcal{U}_\xi(\boldsymbol{y}) = \{\boldsymbol{u} : \ \|\boldsymbol{u}\| = \|\boldsymbol{y}\|, \ \angle(\boldsymbol{y}, \boldsymbol{u}) \le \xi\} \subset \mathcal{B}(\boldsymbol{y}, \delta).$$

$\mathcal{U}_\xi(\boldsymbol{y})$ is compact and w_+ is continuous on this set. Thus

$$\sup_{\mathbf{A} \in \mathbb{A}_\xi} \Big\{ w_+(\mathbf{A}\boldsymbol{y}) f_{\boldsymbol{s}}(\boldsymbol{y}) \Big\} = f_{\boldsymbol{s}}(\boldsymbol{y}) \max \Big\{ w_+(\boldsymbol{u}) : \ \boldsymbol{u} \in \mathcal{U}_\xi(\boldsymbol{y}) \Big\}.$$

Therefor for all $\varepsilon > 0$ there exists angle $\varphi_0 > 0$ such that for all $\xi, \ 0 \le \xi < \varphi_0$, inequality

$$\Big| \sup_{\mathbf{A} \in \mathbb{A}_\xi} \{ w_+(\mathbf{A}\boldsymbol{y}) f_{\boldsymbol{s}}(\boldsymbol{y}) \} - w_+(\boldsymbol{y}) f_{\boldsymbol{s}}(\boldsymbol{y}) \Big| = f_{\boldsymbol{s}}(\boldsymbol{y}) \Big| \max_{\boldsymbol{u} \in \mathcal{U}_\xi(\boldsymbol{y})} w_+(\boldsymbol{u}) - w_+(\boldsymbol{y}) \Big| < \varepsilon$$

holds for all $\boldsymbol{y} \in \mathbb{R}^d \setminus (\partial\mathcal{M} \cup \mathcal{K})$, where $\mathcal{K} = \{\boldsymbol{y} : \ f_{\boldsymbol{s}}(\boldsymbol{y}) = +\infty\}$. Whence condition (C) holds, because Lebesgue measure of $(\partial\mathcal{M} \cup \mathcal{K})$ is equal to zero.

The regularity of infimum function is proved analogically.

□

There is a natural question what can be said about the points outside $\mathcal{H}$ and about the set $\mathcal{H}$ itself. First of all, let us show two counterexamples to the consistency of sample depth (see Figure 2.5).

We consider a uniform distribution on a "hourglass" set, and a uniform distribution on "four tiles". In both cases the distributions are symmetric around a naturally defined central point x and it is exactly the point x where the problem arises. For any sample size n there exists a.s. an orthogonal transformation A such that $\mathrm{E}_n w_+\big(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})\big) = 0$ while $\mathrm{E}_n w_-\big(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})\big) > 0$. In both cases the central point x is *the only* point for which the sample depth is not consistent. Both points are also points of discontinuity of the depth function. Indeed, the theoretical depth $\mathrm{D}(\boldsymbol{x}) = 1$ as follows from the symmetry of distribution. On the other hand there exists sequence $\boldsymbol{x}_n \to \boldsymbol{x}$ such that $\mathrm{D}(\boldsymbol{x}_n) = 0$ for all n.

The nature of the problem lies in the limit of $0/0$ type. Assume without loosing the generality that the central point $\boldsymbol{x} = \mathbf{0}$. In both cases there exists an orthogonal transformation $\mathbf{A}_0$ and a sequence of orthogonal transformations $\mathbf{A}_n$ such that

$$\begin{aligned} &\mathrm{E}w_+\big(\mathbf{A}_0\boldsymbol{X}\big) = 0, \quad \mathrm{E}w_-\big(\mathbf{A}_0\boldsymbol{X}\big) = 0 \\ &\mathrm{E}w_+\big(\mathbf{A}_n\boldsymbol{X}\big) > 0, \quad \mathrm{E}w_-\big(\mathbf{A}_n\boldsymbol{X}\big) > 0 \quad \forall n \\ &\mathrm{E}w_+\big(\mathbf{A}_n\boldsymbol{X}\big) \to 0, \quad \mathrm{E}w_-\big(\mathbf{A}_n\boldsymbol{X}\big) \to 0 \quad \text{as } n \to \infty \end{aligned} \tag{2.17}$$

There exist technical assumptions on the support of probability measure P and on the weight function (beside the regularity conditions of Definition 2.3) such that

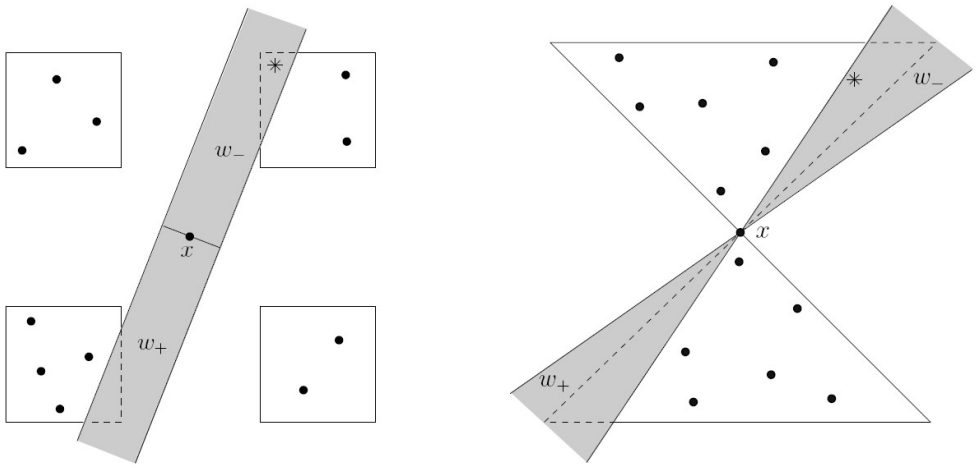


Figure 2.5: The sample depth need not to be consistent.

(2.17) does not hold for any point $\boldsymbol{x} \in \mathbb{R}^d$. Obviously, the critical points are in the interior of convex support and simultaneously in the complement of interior of support itself.

Therefore, if $\mathrm{sp(P)} = \mathrm{csp(P)}$ then $\mathcal{H} = \mathbb{R}^d$ and the strong consistency holds for any point. An example may be normal distribution, bivariate exponential distribution, and many others.

As we have mentioned above, there are technical conditions on the support of probability measure P and on the weight function w_+ such that the consistency hold for $\boldsymbol{y} \in \mathbb{R}^d$. An example of such sufficient conditions may be

- There exist $r > 0$ and $w > 0$ such that $w_+(\boldsymbol{y}) \geq w$ if $y_1^2 + \cdots + y_{d-1}^2 \leq r$.
- There exists a compact set C such that $\mathrm{csp(P)} \setminus \mathrm{sp(P)} \subset C$.
- The interior of support $\mathrm{sp(P)}$ is a connected set.

These conditions are neither necessary conditions, nor the only possible sufficient conditions. In general, the set of points for which the consistency does not hold is, however, small in the sense of probability. Indeed, for any absolutely continuous distribution P it holds

$$\mathrm{P}\{\boldsymbol{y} : \mathrm{D}(\boldsymbol{y}; \mathrm{P}_n) \to \mathrm{D}(\boldsymbol{y}; \mathrm{P}),\ \text{a.s.}\} = 1.$$

The non-consistent points are, as may be clear from the counterexamples, special cases and may be considered as rather "pathological". In particular, consider the "hourglass" distribution together with the band weight function (rather than with the cone weight function) then the consistency of depth holds for the central point $\boldsymbol{x}$ as well as for any other points $\boldsymbol{y} \in \mathbb{R}^2$. Hence, it is a combination of a specific weight function and a specific distribution which causes the trouble at $\boldsymbol{x}$.

2.5 Choice of weight function

In Section 2.4 we discussed the choice of weight function in the Definition 2.1 from the view of consistency of the depth function. Some regularity conditions on

weight function were defined (Definition 2.3 in Section 2.4.1). Further discussion of these conditions was provided in Section 2.4.4. Simple sufficient conditions on the weight function, that lead to the strong consistency of the depth function, were stated in Theorem 2.11. However, another aspects of weight function choice (as discussed in [55]) should be also considered.

An appropriate choice should guarantee that the depth of points lying out of the probability support is equal to zero. Considering any probability measure with a nonconvex support, the requirement might be weaken. Nevertheless, the depth of points lying out of the closed convex support csp(P) of probability measure P should be equal to zero. Following theorem gives a simple condition, which ensures this property.

Theorem 2.12 *Consider the weight function w_+ such that $w_+(\boldsymbol{x}) > 0$ if $x_1^2 + \cdots + x_{p-1}^2 < k$ and $w_+(\boldsymbol{x}) = 0$ elsewhere (k may be infinite). Then $\mathrm{D}(\boldsymbol{x}; \mathrm{P}) = 0$ for any $\boldsymbol{x} \notin \mathrm{csp}(\mathrm{P})$.*

Proof: Note that under the assumptions on w_+ there for all $\boldsymbol{x} \in \mathbb{R}^d$ exists an orthogonal matrix $\mathbf{A}_{\boldsymbol{x}}$ such that $\mathrm{E}w_+\big(\mathbf{A}_{\boldsymbol{x}}(\boldsymbol{X}-\boldsymbol{x})\big) > 0$. It is clear that $\mathrm{D}(\boldsymbol{x}; \mathrm{P}) > 0$ implies that for all orthogonal matrices $\mathbf{A}$ it holds

$$\mathrm{E}w_+\big(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})\big) > 0 \Rightarrow \mathrm{E}w_-\big(\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})\big) = \mathrm{E}w_+\big(\mathbf{I}_-\mathbf{A}(\boldsymbol{X} - \boldsymbol{x})\big) > 0. \quad (2.18)$$

Consider $\boldsymbol{x} \notin \mathrm{csp}(\mathrm{P})$ such that $\mathrm{D}(\boldsymbol{x}; \mathrm{P}) > 0$. It follows from (2.18) that $\boldsymbol{x}$ is "surrounded" by points of sp(P) and therefore $\boldsymbol{x}$ is in the closed convex support of P.

□

On the other hand, a point $\boldsymbol{x} \in \mathrm{int}\big(\mathrm{csp}(\mathrm{P})\big)$ (here $\mathrm{int}(M)$ denotes the interior of a set) need not to be of positive depth. This is a difference from the halfspace depth, since

$$\boldsymbol{x} \in \mathrm{int}\big(\mathrm{csp}(\mathrm{P})\big) \Rightarrow \mathrm{HD}(\boldsymbol{x}) > 0.$$

Indeed, consider uniform distribution on a set

$$S = \{(x, y) : x > 0,\ 1 < x^2 + y^2 < 2\},$$

and a point $\boldsymbol{a} = (x_0, y_0) = (1/2, 0)$. Consider the depth function based on the band weight function of example 1 in section 2.2, where $r^2 < 3/4$. Indeed, for the direction $\boldsymbol{s} = (-1, 0)$ it is clear that $p_+(\boldsymbol{a}, \boldsymbol{s})/p_-(\boldsymbol{a}, \boldsymbol{s}) = 0$ and hence $\mathrm{D}\big((1/2, 0)\big) = 0$. The example is shown in Figure 2.6.

The condition on w_+ given in Theorem 2.12 is of a practical use, but could be weaken. In what follows we tried to find some less restrictive condition on w_+, which ensures depth of points lying out of the closed convex support of distribution equal to zero.

First, consider a probability distribution with a convex support, for example the multivariate exponential distribution or the uniform distribution on a convex support. Let us consider only two-dimensional space ($d = 2$) for simplicity.

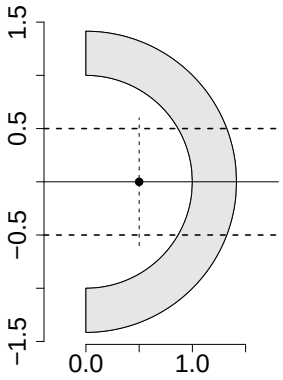


Figure 2.6: Points in closed convex support need not have positive depth.

Theorem 2.13 *Let* P *be a probability measure with a convex support. We denote* $W = \{\boldsymbol{y} : w_+(\boldsymbol{y}) > 0\}$. *Suppose that*

$$\forall \boldsymbol{x} : x_1 = 0, x_2 > 0 \; \exists \; U \text{ a neighbourhood of the point } \boldsymbol{x} : U \subset W, \tag{2.19}$$

holds for the weight function. Then $\mathrm{D}(\boldsymbol{x}; \mathrm{P}) = 0$ *holds for all* $\boldsymbol{x} \notin \mathrm{sp}(\mathrm{P})$.

Proof: Suppose $\boldsymbol{x}_0 \notin \mathrm{sp}(\mathrm{P})$. We want to prove that its weighted depth will be equal to zero when condition (2.19) holds.
We denote $\boldsymbol{x}_m = \arg\min \{|\boldsymbol{x} - \boldsymbol{x}_0| : \boldsymbol{x} \in \mathrm{sp}(\mathrm{P})\}$, i.e. $\boldsymbol{x}_m$ is the point of the support of P with the smallest distance from $\boldsymbol{x}_0$. Existence and uniqueness of this point arise from the convexity of sp(P). We can suppose (without loss of generality) $\boldsymbol{x}_0 = (0, 0)$ because of the translation invariance of the weighted depth (see Theorem 2.3 and Theorem 2.4) and we can suppose (without loss of generality) $\boldsymbol{x}_m = (0, v)$, where $v = |\boldsymbol{x} - \boldsymbol{x}_0|$ because of the rotation invariance of the weighted depth.
For such a rotation $\mathrm{sp}(\mathrm{P}) \subset H_v \subset H_0$ holds, where $H_v = \{\boldsymbol{x} : x_2 \geq v\}$ and $H_0 = \{\boldsymbol{x} : x_2 \geq 0\}$. We can prove it by contradiction. Suppose that there exist $\boldsymbol{y} \in \mathrm{sp}(\mathrm{P})$ such that $y_2 < v$. Then (from the convexity of sp(P)) all points on the abscissa $\boldsymbol{x}_m, \boldsymbol{y}$ are in sp(P), i.e.

$$\boldsymbol{x}_m + \alpha(\boldsymbol{y} - \boldsymbol{x}_m) \; \in \mathrm{sp}(\mathrm{P}) \quad \forall \alpha \in [0, 1].$$

The distance of these points from the origin ($\boldsymbol{x}_0$) can be expressed as $[(\alpha y_1)^2 + (v + \alpha(y_2 - v))^2]^{1/2}$, what is, for alpha small enough, smaller than v. But this is in conflict with the assumption that $\boldsymbol{x}_m = (0, v)$ is the point of sp(P) with the smallest distance from $\boldsymbol{x}_0 = (0, 0)$.
From (2.19) we have that there exist $U_{\boldsymbol{x}_m}$ a neighbourhood of the point $\boldsymbol{x}_m$ such that $U_{\boldsymbol{x}_m} \subset W$. So we have $U_{\boldsymbol{x}_m} \cap W \cap \mathrm{sp}(\mathrm{P}) \neq \varnothing$, hence $\mathrm{E_P}\, w_+ (\boldsymbol{A}\,(\boldsymbol{X} - \boldsymbol{x}_0)) > 0$. It follows from $\mathrm{sp}(\mathrm{P}) \subset H_v \subset H_0$ that $\mathrm{E_P}\, w_- (\boldsymbol{A}\,(\boldsymbol{X} - \boldsymbol{x}_0)) = 0$, so we have $\mathrm{D}(\boldsymbol{x_0}; \mathrm{P}) = 0$.

□

The class of functions that satisfy the condition (2.19) of Theorem 2.13 is still quite broad. For example, all weight functions considered in Section 2.2 satisfy the condition, with the only exception: local weight function (example 3). The impropriety of the local weight function has been already discussed in Section 2.2.

The condition in Theorem 2.13 is not too restrictive. However, this condition is not necessary. For example, consider the uniform distribution on the convex (bounded) support and weight function

$$w_+(\boldsymbol{x}) > 0 \text{ if and only if } x_d \in [0, h],$$

where h is a positive constant. The condition (2.19) is not satisfied, but $\boldsymbol{x} \notin \mathrm{csp(P)} \Rightarrow \mathrm{D}(\boldsymbol{x}; \mathrm{P}) = 0$. holds.

The form of the condition (2.19) for general d-dimensional case can be written as follows: For all $\boldsymbol{x} : x_1 = 0, \ldots, x_{d-1} = 0, x_d > 0$ there exist U, a neighbourhood of the point $\boldsymbol{x}$, such that $U \subset W$.

Now we will consider a probability measure P with nonconvex support. Because of nonconvexity we do not request depth equal to zero for all points out of the support itself, but only for all points out of the closed convex hull of the support ($\mathsf{csp}(\mathrm{P})$). For example, if we consider the uniform distribution on an annulus, the center of this annulus is the center of symmetry of the distribution, but it is still out of the support. So we do not demand its depth equal to zero.

For a nonconvex support of P the condition (2.19) is no more tenable. We can either consider $\mathsf{sp}(\mathrm{P})$ connected or strengthen the condition on weight function. One possible way how to do it for such a case is introduced in Theorem 2.14.

Theorem 2.14 *Suppose there exists $n \in \mathbb{N}$ such that $\mathsf{sp}(\mathrm{P}) = \bigcup_{i=1}^{n} K_i$, where K_i (i=1,...,n) is connected subset of $\mathbb{R}^2$, and $\mathsf{sp}(\mathrm{P})$ has no singular point.*
Denote $m_{ij} = \min\{|\boldsymbol{x} - \boldsymbol{y}| : \boldsymbol{x} \in K_i, \boldsymbol{y} \in K_j\}$, $i, j = 1, \ldots, n$.
Consider $m = \max_{1 \le i,j \le n} m_{ij}$.
Let a weight function w_+ have the following property:

$$\forall \boldsymbol{x} : |x_1| \le m/2 \ \exists\ U_{\boldsymbol{x}} \text{ a neighbourhood of } \boldsymbol{x} \text{ such that } U_{\boldsymbol{x}} \subset W. \tag{2.20}$$

Then $\boldsymbol{x} \notin \mathsf{csp}(\mathrm{P}) \Rightarrow \mathrm{D}(\boldsymbol{x}; \mathrm{P}) = 0$.

Proof: The proof is very similar to the previous one. We denote the point of csp(P) with the smallest distance from $\boldsymbol{x}_0 = (0, 0)$ by $\boldsymbol{x}_m = (0, v)$.
We will prove that there exist a point $\boldsymbol{x} = (x_1, x_2)$ such that $|x_1| \le m/2$ which is in W ($\boldsymbol{x} \in W$) by contradiction. Suppose that there is no such a point $\boldsymbol{x}$. Then there must be points $\boldsymbol{y} = (y_1, y_2) \in \mathsf{sp}(P)$ and $\boldsymbol{z} = (z_1, z_2) \in \mathsf{sp}(P)$ such that $y_1 < -m/2$ and $z_1 > m/2$. Hence $|\boldsymbol{y} - \boldsymbol{z}| > m$. These points are from the different components of connectedness. We take two points from these components with the smallest distance between each other: $\boldsymbol{y}_m$ and $\boldsymbol{z}_m$. For all points $\boldsymbol{y}$ of the one component $y_1 < -m/2$ holds and for all points $\boldsymbol{z}$ of the other component $z_1 > m/2$ holds, we get $|\boldsymbol{y}_m - \boldsymbol{z}_m| > m$, but this is in conflict with the definition of m.
From the assumption (2.20) follows that there exists $U_{\boldsymbol{x}}$ a neighbourhood of the point $\boldsymbol{x}$ such that $U_{\boldsymbol{x}} \cap W \cap \mathsf{sp}(\mathrm{P}) \neq \varnothing$. Hence (similarly as in proof of the Theorem 2.13) $\mathrm{D}(\boldsymbol{x}; \mathrm{P}) = 0$.

□

Note that in a special case $n = 1$ (for nonconvex connected support) we have depth equal to zero for all points out of the closed convex hull of the support when (2.19) holds.

We have been discussing the depth of the points out of the closed convex hull of the support so far. Now we will discuss properties of the depth of points that are in the convex hull of the support, but out of the support itself. It is easy to show that all points from the closed convex hull of the support have a positive halfspace depth. An advantage of the weighted halfspace depth is that points in the closed convex hull of the support but out of the support itself might have the depth equal to zero.

We explain the advantage on the following example. Consider the uniform distribution on some sector which originates from the circle with the radius r (see Figure2.7, part a). All points from the closed convex hull of the support have the halfspace depth greater then zero (Figure 2.7, part b). Now we consider the weighted halfspace depth with the following weight function

$$\begin{aligned} w_+(\boldsymbol{x}) &= 1 \quad \text{if } |x_1| < h,\ x_2 \geq 0 \\ &= 0 \quad \text{otherwise,} \end{aligned} \tag{2.21}$$

where h is some positive constant smaller than r (the band depth from example 1 at the end of the Section2.1). The area of points that have the weighted halfspace depth greater then zero is the union of the support and the circle with the same center as the big one, but with radius equal to h (Figure 2.7, part c). Comparing shapes of the areas with nonzero halfspace depth and nonzero weighted halfspace depth we see that the second one is more similar to the shape of the support of probability measure.

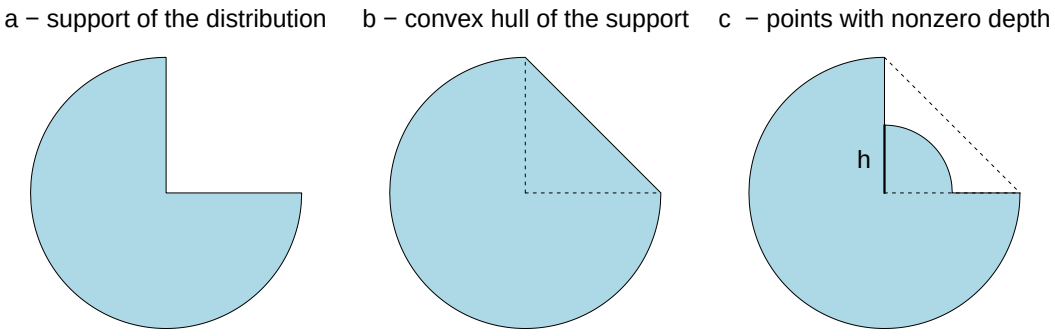


Figure 2.7: Sector-shaped support of the uniform distribution (a); the convex hull of the support i.e. points with nonzero halfspace depth (b); points with nonzero weighted halfspace depth with the weight function (2.21), where $h = r/2$ (c).

We illustrate the example with the results of simulation study. We generated 1000 points from the uniform distribution of the considered sector-shaped support. Figure 2.8 show big differences between the areas of deepest points when using halfspace and weighted halfspace depth.

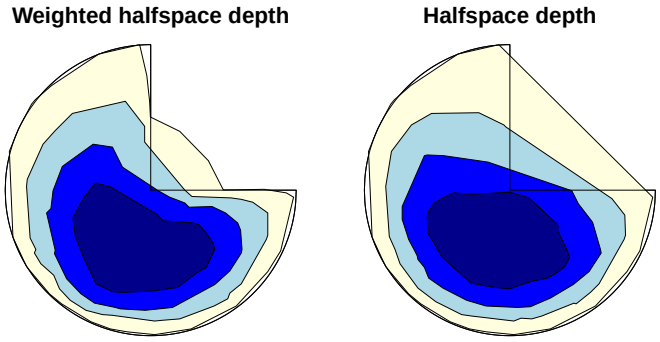


Figure 2.8: Uniform distribution on the sector-shaped support: areas of 25%, 50% and 75% of deepest points.

2.6 Computational aspects

In this section, we shortly discuss the computational aspects of sample depth computation.

Since the weighted halfspace depth is defined for a broad class of weight functions, a general fast algorithm for depth computing does not exist. Also, the theoretical depth $D(\boldsymbol{x}; \mathrm{P})$ of point $\boldsymbol{x}$ under a general absolutely continuous distribution P cannot be usually calculated exactly and some numerical approximation is needed. It is caused by the fact that $w_+(\mathbf{A}\boldsymbol{x})$ can attain different values for every transformation $\mathbf{A} \in \mathbb{O}_d$, which means that possibly uncountable number of values must be considered. The symmetric weight functions allow to use only rotation rather than all orthogonal transformations $\mathbb{O}_d$.

On the other hand, in some special cases the empirical depth may be computed exactly. It is the case when the weight function is piecewise constant. The cone weighted depth, the band weighted depth, the halfspace depth are, in particular, examples of such depths. The set $\{\sum_{i=1}^{n} w_+\mathbf{A}(\boldsymbol{X}_i - \boldsymbol{x}), \mathbf{A} \in \mathbb{O}_d\}$ is finite for each $\boldsymbol{x}$ in such a case.

Straightforward algorithm is used to compute the sample depth of a given point $\boldsymbol{x}$. It uses a predefined number of vectors in $\mathbb{R}^{d-1} \times [0, +\infty)$ which represent halfspaces in which we compute sample weighted probability. These vectors are normal vectors of hyperplanes which determine appropriate halfspaces. For every such vector we rotate our dataset so that normal vector goes to x_d axis. Then we make, for rotated dataset, two computations of sample weighted probability - for halfspace where $x_d \geq 0$ and for halfspace where $x_d \leq 0$. Finally the depth is set to the smallest value of portions of sample weighted probabilities in $x_d \geq 0$ and $x_d \leq 0$ halfspaces. For sample size n the computation of weighted probability in given halfspace takes $O(n)$ steps. There are $2k$ halfspaces, hence computation of depth of given point takes $O(2kn)$ steps. If one wants to compute the depth of all points in dataset it takes $O(2kn^2)$ steps. Note that for two dimensional dataset setting the choice $2k = 1000$ halfspaces brings very precise answer.

2.7 Examples

We illustrate some differences between the weighted depth and the halfspace depth. In the following examples we use the band weight function of example 1 in Section 2.2. We call this depth as the band weighted depth or simply about the *band depth*. Several bivariate distributions of random vector $\boldsymbol{X} = (X_1, X_2)^T$ are considered.

We simulate 2500 points for each particular distribution and we compute sample depth of these points. In next figures the areas of 25%, 50% and 75% of the deepest points (points with the highest depth) are plotted. The rest of points (25% points with the lowest depth) are marked by light grey. A triangle marks the sample deepest point.

2.7.1 Symmetrical distributions with convex levelsets of density

First let us consider two cases with natural centre - normal distribution $\mathsf{N}_2(\mathbf{0}, \mathbf{I}_2)$ and uniform distribution on the unit square $[-0.5, 0.5] \times [-0.5, 0.5]$. We can compute the depth of any point exactly in the first case.

Let $\boldsymbol{X}$ be a two dimensional random vector with normal distribution $\mathsf{N}_2(\mathbf{0}, \mathbf{I}_2)$. Suppose we have a band weight function

$$w_+(x_1, x_2) = \begin{cases} 1, & \text{if } -h < x_1 < h,\ x_2 > 0 \\ 0, & \text{otherwise} \end{cases}$$

for given $h > 0$. We use the same notation as in example 1 of section 2.2, hence

$$\mathrm{D}(\boldsymbol{x}) = \inf_{\|\boldsymbol{s}\|=1} \frac{p_+(\boldsymbol{x}, \boldsymbol{s})}{p_-(\boldsymbol{x}, \boldsymbol{s})}. \tag{2.22}$$

First we show that for an arbitrary point $\boldsymbol{x}$ it holds

$$\mathrm{D}(\boldsymbol{x}) = \min\left\{ \frac{p_+(\boldsymbol{x}, \boldsymbol{s}_0)}{p_-(\boldsymbol{x}, \boldsymbol{s}_0)}, \frac{p_-(\boldsymbol{x}, \boldsymbol{s}_0)}{p_+(\boldsymbol{x}, \boldsymbol{s}_0)} \right\}$$

for $\boldsymbol{s}_0$ such that $\mathbf{0} \in \{\boldsymbol{x} + t\boldsymbol{s}_0,\ t \in \mathbb{R}\}$.

Without loss of generality we can assume that $\boldsymbol{x} = (0, x_2)^T$ (the distribution is symmetric about $\mathbf{0}$ and also about any line containing $\mathbf{0}$). For such a point $\boldsymbol{x}$ let $\boldsymbol{s} = (0, 1)^T$. One has

$$\begin{aligned} p_+(\boldsymbol{x}, (0,1)^T) &= \mathrm{P}(X_2 > x_2,\ -h < X_1 < h) = (1 - \Phi(x_2))\mathrm{P}(-h < X_1 < h), \\ p_-(\boldsymbol{x}, (0,1)^T) &= \mathrm{P}(X_2 < x_2,\ -h < X_1 < h) = \Phi(x_2)\mathrm{P}(-h < X_1 < h), \end{aligned}$$

where Φ is the distribution function of $\mathsf{N}(0, 1)$. For any other direction $\boldsymbol{u} \neq \boldsymbol{s}$ and bands $\mathbb{B}(\boldsymbol{x}, \boldsymbol{u})$ there exists uniquely determined rotation $\mathbf{A} \in \mathbb{O}_2$ such that $\mathbf{A}\boldsymbol{u} = (0, 1)^T$ and $\mathbf{A}\boldsymbol{X} = \boldsymbol{X}' \sim \mathsf{N}_2(\mathbf{0}, \mathbf{I}_2)$. For $\boldsymbol{x} = (0, x_2)^T$ it holds $\mathbf{A}\boldsymbol{x} = \boldsymbol{x}'$ where $x_2 > x_2'$. It is easy to show that

$$\begin{aligned} p_+(\boldsymbol{x}, \boldsymbol{u}) &= p_+(\boldsymbol{x}', (0,1)^T) = \mathrm{P}(X_2' \geq x_2')\mathrm{P}(x_1' - h < X_1' < x_1' + h) \\ &= (1 - \Phi(x_2'))\mathrm{P}(x_1' - h < X_1' < x_1' + h), \\ p_-(\boldsymbol{x}, \boldsymbol{u}) &= \Phi(x_2')\mathrm{P}(x_1' - h < X_1' < x_1' + h). \end{aligned}$$

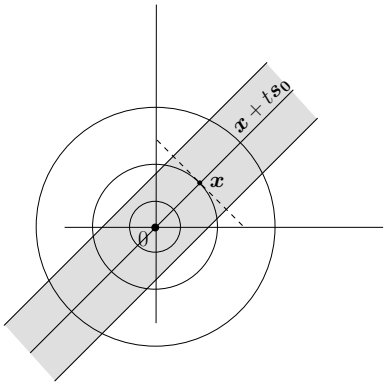


Figure 2.9: The direction, in which the ratio of weighted halfspaces is minimal, is determined by $\boldsymbol{x}$ and $\mathbf{0}$.

Since $\Phi(x_2) > \Phi(x_2')$ it follows

$$\frac{p_+(\boldsymbol{x},\boldsymbol{u})}{p_-(\boldsymbol{x},\boldsymbol{u})} = \frac{1-\Phi(x_2')}{\Phi(x_2')} > \frac{1-\Phi(x_2)}{\Phi(x_2)} = \frac{p_+(\boldsymbol{x},(0,1)^T)}{p_-(\boldsymbol{x},(0,1)^T)}.$$

Hence

$$\mathrm{D}(\boldsymbol{x}) = \frac{1-\Phi(x_2)}{\Phi(x_2)}.$$

Since both the depth function and the distribution are invariant with respect to rotation, it follows that for any $\boldsymbol{y} \in \mathbb{R}^2$

$$\mathrm{D}(\boldsymbol{y}) = \mathrm{D}\big((0,\|\boldsymbol{y}\|)^T\big) = \frac{1-\Phi(\|\boldsymbol{y}\|)}{\Phi(\|\boldsymbol{y}\|)}.$$

The depth does not depend on the value of h and it is equal to the halfspace depth.

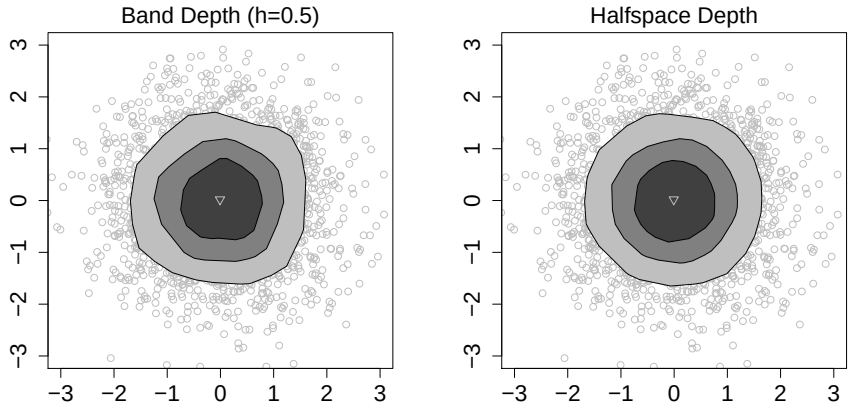


Figure 2.10: Normal distribution $\mathsf{N}_2(\mathbf{0},\mathbf{I}_2)$: areas of 25%, 50% and 75% of deepest points.

In Figure 2.10 we can see the sample areas of the deepest points based on simulation. There is no big difference between the band weighted depth and the halfspace depth for bivariate normal distribution $\mathsf{N}_2(\mathbf{0},\mathbf{I}_2)$. Both methods find point $(-0.008, 0.019)$ as the sample deepest point, which is the "observation" (sample point) with the smallest distance (in the standard Euclidean metric) from

the theoretical centre $(0, 0)$. Areas of the deepest points are similarly large. The only remarkable difference is in the value of sample depth in sample deepest point (recall that this is the same point for both methods) which is 0.88 for the band depth and 0.94 corresponding to the halfspace depth (theoretical depth of the deepest point is equal to one in both cases). Differences between the sample band weighted depth and the sample halfspace depth for fixed sample size become smaller as h increases. These results are in accordance with the theoretical result above.

Consider now the case of uniform distribution on the square support, for example on $[-0.5, 0.5] \times [-0.5, 0.5]$. We want to calculate the halfspace depth of a point $\boldsymbol{x} = (x_1, x_2)^T$. Without loss of generality we can assume $x_1 \geq 0$, $x_2 \geq 0$ and $x_2 \geq x_1$ (otherwise symmetry can be used). Each line going through the point $\boldsymbol{x}$ divides the square into two parts. The halfspace depth of $\boldsymbol{x}$ is equal to the minimal volume of the smaller part, when all possible lines going through $\boldsymbol{x}$ are considered. It is easy to show that such a line have to intersect right and upper side of the square. The situation is shown in left part of Figure 2.11.

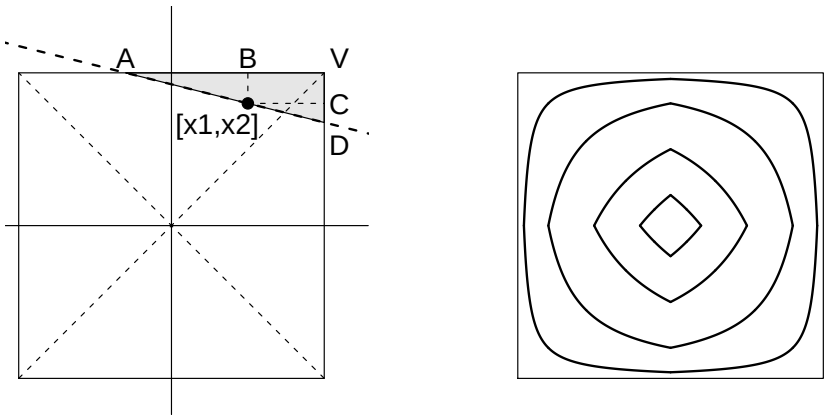


Figure 2.11: Computing the halfspace depth of a point $\boldsymbol{x} = (x_1, x_2)^T$ (left) and contours of the central areas of uniform distribution on square support according to the halfspace depth (right).

The volume of gray area can be expressed as:

$$S = (|AB| + |BV|) \cdot (|CV| + |CD|) /2,$$

where

$$\begin{aligned} |BV| &= 0.5 - x_1, \\ |CV| &= 0.5 - x_2, \\ |AB| &= \frac{|BV| \cdot |CV|}{|CD|}. \end{aligned}$$

The last equality is from the similarity of triangles ABX and AVD. Using equalities above and minimizing S as a function of $|CD|$ leads to the equality $|AB| = |BV|$ (and thus $|CV| = |CD|$). In this case the volume S is minimal and

it is equal to $2(0.5 - x_1)(0.5 - x_2)$. Thus we have for all $\boldsymbol{x} = (x_1, x_2)^T$, such that $x_1 \geq 0$, $x_2 \geq 0$ and $x_2 \geq x_1$:

$$D(\boldsymbol{x}) = 2 \cdot (0.5 - x_1) \cdot (0.5 - x_2).$$

Using symmetry we get a formula for the halfspace depth on any $\boldsymbol{x} \in [-0.5, 0.5] \times [-0.5, 0.5]$:

$$D(\boldsymbol{x}) = 2 \cdot \min(|0.5 - x_1|, |0.5 + x_1|) \cdot \min(|0.5 - x_2|, |0.5 + x_2|).$$

The deepest point is the centre of symmetry with halfspace depth equal to 1/2; all points at the border of the support have the halfspace depth equal to zero. The contours of central areas are compound from hyperbolic parts as can be seen at the right side of the Figure2.11. The shape of central areas is rather disappointing - it does not reflect the square shape of the support.

Consider now the band depth function. The shape of central areas is determined by the band width parameter h. Therefore we will denote the band depth function as $D_h(\cdot)$. If $h \to \infty$, the depth is equivalent to the halfspace depth and the central areas are as was shown above. On the other hand, consider the situation $h \to 0$. In this case the depth of any point $\boldsymbol{x} \in \text{int}(\mathsf{sp}(P))$ can be compute quite easily. The following theorem deals with the limit case.

Theorem 2.15 *Let P be a uniform distribution on a convex (bounded) centrally symmetric support in $\mathbb{R}^d$. For any $\boldsymbol{x} \in \text{int}(\mathsf{sp}(P))$ denote $\mathcal{L}(\boldsymbol{x})$ the set of all lines containing the point $\boldsymbol{x}$ and define*

$$D_0(\boldsymbol{x}) = \inf_{l \in \mathcal{L}(\boldsymbol{x})} \frac{\|\boldsymbol{x} - H_1(l)\|}{\|\boldsymbol{x} - H_2(l)\|},$$

where $H_1(l)$ and $H_2(l)$ are the intersects of the line l and the boundary of $\mathsf{sp}(P)$.

Then for any $\boldsymbol{x} \in \text{int}(\mathsf{sp}(P))$ it holds:

$$D_h(\boldsymbol{x}) \to D_0(\boldsymbol{x}), \quad as \quad h \to 0.$$

Denote the boundary of $\mathsf{sp}(P)$ by B and the centre of symmetry by S. Following lemma lighten the proof:

Lemma 2.16 *Consider the situation described in Theorem 2.15. Then for any $\boldsymbol{x} \in \text{int}(\mathsf{sp}(P))$ it holds*

$$D_0(\boldsymbol{x}) = \frac{\|\boldsymbol{x} - H_1(l_0)\|}{\|\boldsymbol{x} - H_2(l_0)\|},$$

where l_0 is the line going through the points $\boldsymbol{x}$ and S, $H_1(l_0)$ denotes the intersect of boundary B with the half-line from S to $\boldsymbol{x}$, and $H_2(l_0)$ denotes the intersect of boundary B and half-line from $\boldsymbol{x}$ to S.

Proof: Suppose for the sake of contradiction that there is another $l_1 \in \mathcal{L}$ ($l_1 \neq l_0$), such that

$$\frac{\|\boldsymbol{x} - H_1(l_1)\|}{\|\boldsymbol{x} - H_2(l_1)\|} < \frac{\|\boldsymbol{x} - H_1(l_0)\|}{\|\boldsymbol{x} - H_2(l_0)\|} \tag{2.23}$$

Consider now the plane determined by l_0 and l_1. The triangle $H_2(l_0)H_1(l_0)H_2(l_1) \subset \mathsf{sp}(P)$ because of convexity of the $\mathsf{sp}(P)$. Let $H_2^*(l_1)$ denote the point, which

is symmetrical to $H_2(l_1)$ around the centre of symmetry S. Then the triangle $H_2(l_0)H_1(l_0)H_2^*(l_1)$ is also a subset of $\mathsf{sp}(P)$ (from convexity assumption). Denote the intersect of l_1 and $H_2^*(l_1)H_1(l_0)$ by H. Then $H_1(l_1)$ lies on the half-line opposite to $HH_2(l_1)$. Hence

$$\frac{\|\boldsymbol{x} - H_1(l_1)\|}{\|\boldsymbol{x} - H_2(l_1)\|} \geq \frac{\|\boldsymbol{x} - H\|}{\|\boldsymbol{x} - H_2(l_1)\|} = \frac{\|\boldsymbol{x} - H_1(l_0)\|}{\|\boldsymbol{x} - H_2(l_0)\|}.$$

The last equality follows from the similarity of triangles $H_2(l_0)\boldsymbol{x}H_2(l_1)$ and $H_1(l_0)\boldsymbol{x}H$. This contradicts to inequality (2.23).

□

Proof of the Theorem 2.15: We show the proof for two-dimensional case. Without loss of generality we can assume $\mathsf{sp}(P)$ having volume equal to one. For a particular $\boldsymbol{x} \in \mathrm{int}(\mathsf{sp}(P))$ denote:

$$\begin{aligned} r &= \min_{y \in B} \|x - y\| \;\; > \;\; 0, \\ R &= \max_{y \in B} \|x - y\| \;\; > \;\; 0. \end{aligned}$$

It is sufficient to prove that

$$\forall \epsilon > 0 \; \exists h_0 > 0 \; \forall l \in \mathcal{L}(\boldsymbol{x}) : \; h < h_0 \Rightarrow \left| \frac{p_+(\boldsymbol{x}, l)}{p_-(\boldsymbol{x}, l)} - \frac{\|\boldsymbol{x} - H_1(l)\|}{\|\boldsymbol{x} - H_2(l)\|} \right| < \epsilon.$$

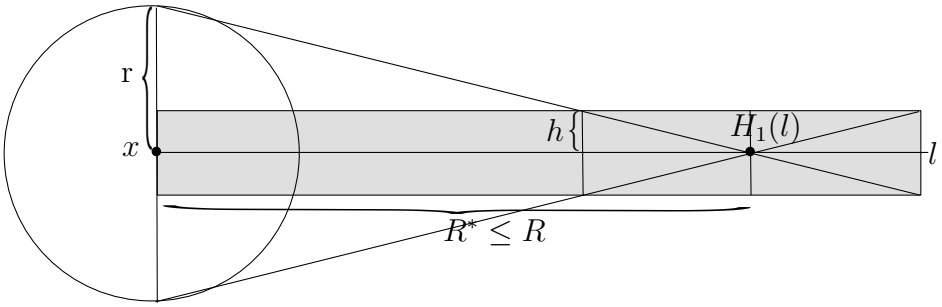


Figure 2.12: Scheme of maximal possible difference between $p_+(\boldsymbol{x}, l)$ and a product $2h \cdot \|\boldsymbol{x} - H_1(l)\|$.

Let $h < r$. Then it holds:

$$\begin{aligned} |p_+(\boldsymbol{x}, l) - 2h \cdot \|\boldsymbol{x} - H_1(l)\|| < 2h^2 R/r, \\ |p_-(\boldsymbol{x}, l) - 2h \cdot \|\boldsymbol{x} - H_2(l)\|| < 2h^2 R/r. \end{aligned}$$

We can thus write:

$$\begin{aligned} p_+(\boldsymbol{x}, l) &= 2h \cdot \|\boldsymbol{x} - H_1(l)\| + \epsilon_1(l), \quad \text{where } |\epsilon_1(l)| < 2h^2 R/r, \\ p_-(\boldsymbol{x}, l) &= 2h \cdot \|\boldsymbol{x} - H_2(l)\| + \epsilon_2(l), \quad \text{where } |\epsilon_2(l)| < 2h^2 R/r, \end{aligned}$$

and the considered difference can be expressed as

$$\frac{p_+(\boldsymbol{x}, l)}{p_-(\boldsymbol{x}, l)} - \frac{\|\boldsymbol{x} - H_1(l)\|}{\|\boldsymbol{x} - H_2(l)\|} = \frac{\epsilon_1(l) \, \|\boldsymbol{x} - H_2(l)\| + \epsilon_2(l) \, \|\boldsymbol{x} - H_1(l)\|}{2h \cdot \|\boldsymbol{x} - H_2(l)\| + \epsilon_2(l)}. \tag{2.24}$$

For $h < r^2/R$ the denominator is always positive, as $2h \cdot \|\boldsymbol{x} - H_2(l)\| + \epsilon_2(l) \geq$ $\geq 2hr - 2h^2R/r = 2h(r - 2hR/r)$.

From (2.24) we have

$$\left|\frac{p_+(\boldsymbol{x}, l)}{p_-(\boldsymbol{x}, l)} - \frac{\|\boldsymbol{x} - H_1(l)\|}{\|\boldsymbol{x} - H_2(l)\|}\right| \leq \frac{4R^2h^2/r}{2hr - 2Rh^2/r} = \frac{2R^2h}{r^2 - Rh} < \epsilon \quad \text{if } h < r^2\epsilon/[R(2R+\epsilon)].$$

□

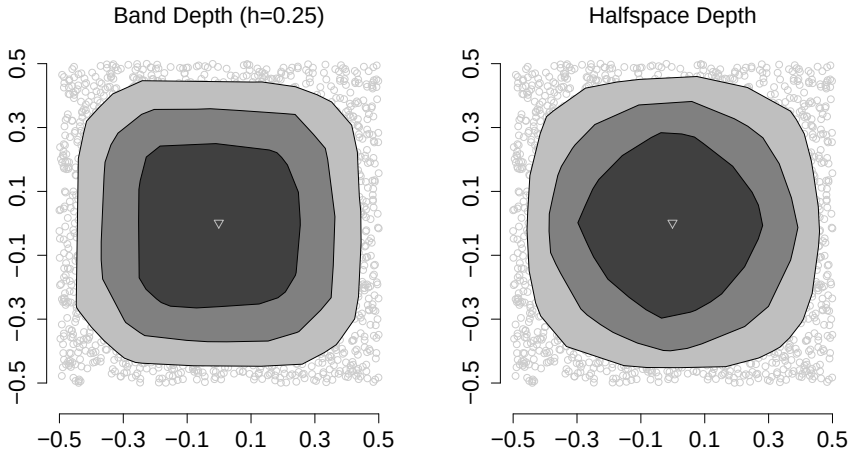


Figure 2.13: Uniform distribution on $[-0.5, 0.5] \times [-0.5, 0.5]$: areas of 25%, 50% and 75% of deepest points.

In Figure 2.13 areas of deepest points for uniform distribution on square $[-0.5, 0.5] \times [-0.5, 0.5]$ are displayed. The difference between band weighted depth (with $h = 0.25$, hence the band width is 0.5) and halfspace depth is obvious. The main difference is in the shape of areas of deepest points. Band depth keeps more faithfully the shape of the support i.e. square whereas halfspace depth areas are rather going to be a circle. It is not a surprise that for uniform distribution the areas are similarly large for both methods and there is a common sample deepest point $(-0.001, 0.002)$, which is pretty close to theoretical centre $(0, 0)$. Again the sample depth of the sample deepest point is remarkably smaller for band depth (0.91 for band depth, 0.96 for twice the halfspace depth).

We should note that differences are going to be smaller and smaller as h increases for both normal and uniform distribution. In the case of uniform distribution there is even no difference between theoretical band depth and halfspace depth if h is greater than the diagonal of the square ($h > \sqrt{2}$).

2.7.2 Symmetrical distributions - a nonconvex case

Centrally symmetric distributions need not have convex levelsets of density. For simplicity, we show here two cases of uniform distribution on nonconvex support. It is easy to imagine distributions with levelsets of density, which have these shapes.

Let us consider the uniform distribution, whose support has a shape of a cross. The difference between central areas of the halfspace depth and the band depth (with h=0.5) estimated by the simulation is shown in Figure 2.14. The central

areas of the halfspace depth are convex. Their estimated version is smoother than the estimated version of the band depth central regions. However, they include also points that are out of the support of the distribution. In contrary, central areas of the band depth reflects better the shape of the support.

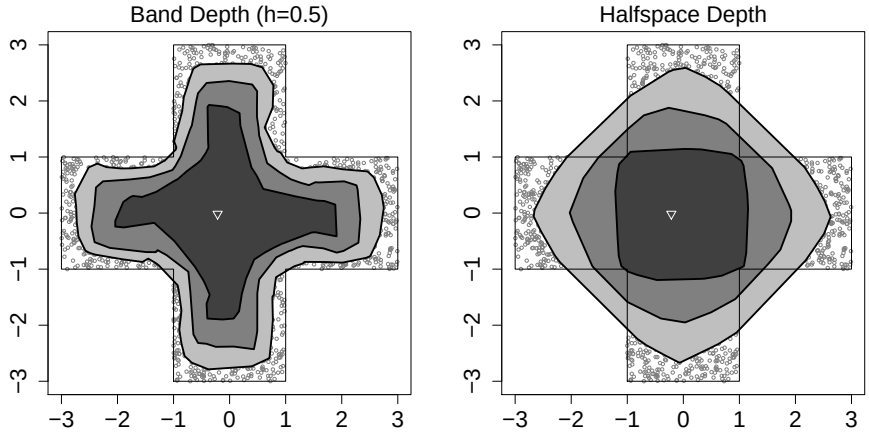


Figure 2.14: Uniform distribution on the cross-shaped support: areas of 25%, 50% and 75% of deepest points.

Another example is a uniform distribution on the support, whose shape consists of all points $(x_1, x_2) \in \mathbb{R}^2$ such that $x_1^2 + x_2^2 \leq 3^2$ and $|x_2/x_1| \leq 1$. The difference between central areas of the halfspace and band depth are shown in Figure 2.15. The main difference is in the convexity of the central areas. Central areas of the band depth again reflects better the shape of the support. The deepest point is located near the centre of symmetry in both cases (however, different points were marked as the deepest points) The halfspace depth of the deepest point is 0.92, while the band depth of the deepest point is only 0.75.

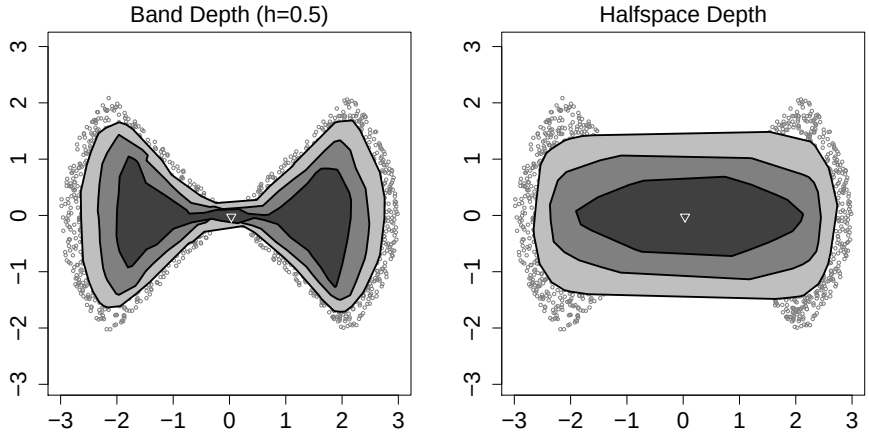


Figure 2.15: Uniform distribution on the noncovex support: areas of 25%, 50% and 75% of deepest points.

2.7.3 Non-symmetrical distributions with convex levelsets of density

In two previous examples we have considered centrally symmetric distributions. For such a distribution there is a naturally defined unique centre. Now we will consider some distributions that are not symmetric and the notion of *centre* may be questionable.

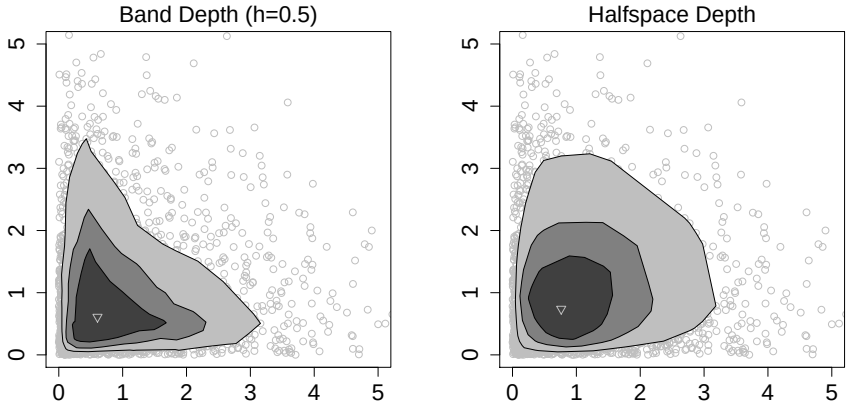


Figure 2.16: Exponential distribution ($X_1 \sim \mathsf{Exp}(1),\ X_2 \sim \mathsf{Exp}(1)$, X_1 and X_2 are independent) : areas of 25%, 50% and 75% of deepest points.

In Figure 2.16 big differences between the band weighted depth and the halfspace depth for exponential distribution can be easily seen. Band depth areas are rather triangular whereas the halfspace depth areas are rather oval. Note that band depth areas correspond better to level sets of the density (level sets of this distribution are rectangular isosceles triangles with vertex in $(0, 0)$). Also there is a remarkable difference in position of sample deepest point which is $(0.606, 0.610)$ for band depth (depth = 0.68) and $(0.763, 0.739)$ for halfspace depth (twice depth = 0.77). Both are close to line $y = x$, but sample deepest point for band depth is closer to 0. Another difference is that areas for the halfspace depth are about 30%-40% larger than for band depth. This can be seen from the Table 2.1.

Proportion of deepest points	**Volume of the area**	
	Halfspace Depth	**Band Depth (h=0.5)**
75%	7.8	5.7
50%	3.4	2.6
25%	1.4	1.0

Table 2.1: Volume of central areas for distribution $X_1 \sim \mathsf{Exp}(1),\ X_2 \sim \mathsf{Exp}(1)$, X_1 and X_2 are independent.

Another example is a random vector (X_1, X_2) where $X_1 \sim \mathrm{Exp}(1)$ and $X_2|X_1 = x \sim N(x, x)$. Both normal and exponential distributions have maximal density for x near zero (we mean right neighbourhood of zero in the case of exponential distribution). In Figure 2.17 we can see that while band depth tends to approach zero, there is a notable lack between zero and area of the deepest points for halfspace depth. We can also observe that areas for band depth have smaller volume than corresponding central areas for halfspace depth, see Table 2.2.

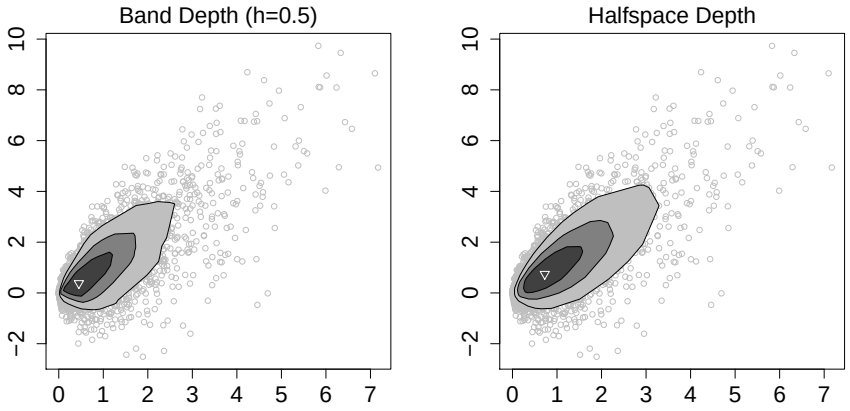


Figure 2.17: $X_1 \sim \text{Exp}(1)$ and $X_2|X_1 = x \sim N(x, x)$: areas of 25%, 50% and 75% of deepest points.

Proportion of deepest points	**Volume of the area**	
	Halfspace Depth	**Band Depth (h=0.5)**
75%	8.7	6.5
50%	3.6	2.6
25%	1.2	0.9

Table 2.2: Volume of central areas for distribution $X_1 \sim \text{Exp}(1)$, $X_2|X_1 = x \sim N(x, x)$.

2.7.4 Non-symmetrical distributions - a nonconvex case

Sometimes the levelsets of the density function of a distribution are not convex. Even the support of the distribution can be nonconvex. Here we show two examples - a mixture of two normal distributions and a uniform distribution on the nonconvex support.

In Figure 2.18 a mixture of two bivariate normal distributions is plotted, namely we considered

$$\mathsf{N}_2\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 & -0.9\sqrt{3} \\ -0.9\sqrt{3} & 1 \end{pmatrix}\right) \text{ and } \mathsf{N}_2\left(\begin{pmatrix} -2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 0.8\sqrt{2} \\ 0.8\sqrt{2} & 1 \end{pmatrix}\right).$$

A remarkable difference between the halfspace depth and the band weighted depth can be seen. Band weighted depth areas again correspond more faithfully to level sets of density. The shape of the areas for band depth give evidence that the distribution is mixture of two other distributions. Areas for halfspace depth are about 25% larger than for band depth. The difference in position of sample deepest point is not surprising. In such a situation (we have two distinct natural centres) estimator of the deepest point for band depth may be quite unstable, because in such cases there need not exist unique deepest point. For band depth the sample deepest point is $(0.099, 0.538)$ (depth = 0.60), for halfspace depth it is $(-0.534, 0.958)$ (twice depth = 0.70). Both these points are quite close to an abscissa that connects theoretical centres of normal distributions (these centres are marked by light circle).

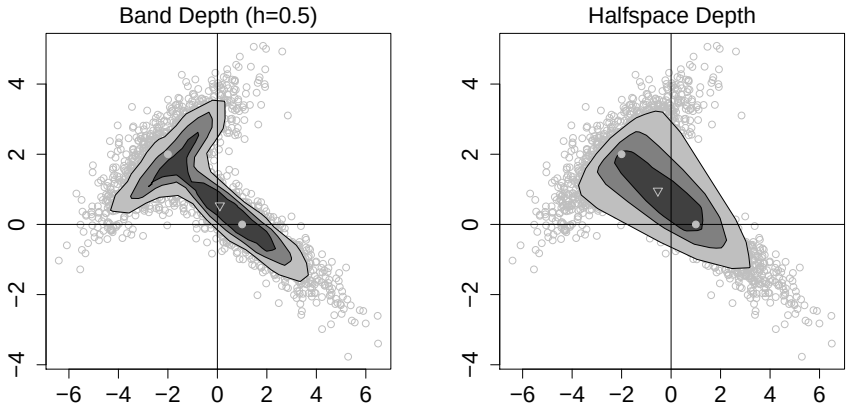


Figure 2.18: Mixture of two bivariate normal distributions: areas of 25%, 50% and 75% of deepest points.

Next example deals with a uniform distribution on a nonconvex support. Figure 2.19 shows that the band depth keeps more faithfully the shape of the support. Central areas of the halfspace depth are always convex sets - we drew only deepest points from the sample. Central areas of the band depth are not convex in this case.

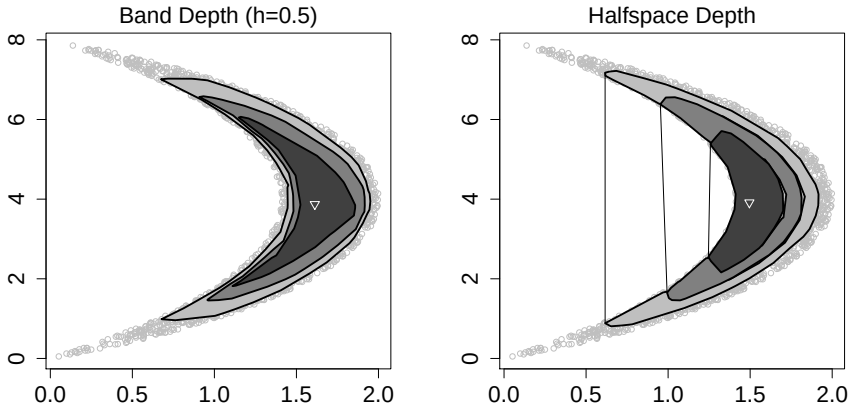


Figure 2.19: Uniform distribution on a nonconvex support: areas of 25%, 50% and 75% of deepest points.

2.7.5 Concluding remarks

Concluding this section we should note that

- Main differences between the band and the halfspace depth are in the shape of areas of deepest points.
- For considered nonsymmetric distributions the areas for the halfspace depth were remarkably larger than for the band depth.
- For symmetric distribution both depths localise the centre of symmetry quite well, for nonsymmetric distributions there are differences in localisation of the deepest point (which may not be unique for the band depth).

Chapter 3

Discrimination based on data depth

3.1 Discrimination problem

Discrimination problem consists in creating a rule for distinguishing objects of several groups. Several objects of our interest form two or more groups according to some criterion (*e.g. people divided by their illness status: suffering and not suffering from a certain disease, series of products divided according to their quality: first class, second class and third class quality*). Division of objects into particular groups was done in the way, which is unknown to us. We know only the final division. Besides the group assignment we know several numerical characteristics of each object of our interest (*e.g. body temperature, blood pressure and number of days with a certain symptom*). Available objects, represented both by their numerical characteristics and group assignment, form so called training set.

For any new object of interest (*e.g. a new patient or a new series of products*) it is possible to measure or observe its numerical characteristics, but not the group, to which it belongs. Our goal is to find a rule, which allocate the new observation to the group, to which it belongs, based only on known characteristics of the new object; more precisely, on their similarity to the characteristics of the objects in particular groups (*e.g. the body temperature of people suffering from influenza is generally higher than the body temperature of healthy people. Thus the person, whose body temperature is* $39°C$*, is more likely to suffer from influenza than a person with usual body temperature of* $36.5°C$). Such a rule is called classifier. The goal of the discriminant analysis is to find some classifier based on available records of previously classified objects.

It should be noticed that sometimes the term *classification* is used instead of *discrimination*. We keep the terminology in accordance with Hand [19], who distinguish the *discrimination* as the process of deriving classification rules from samples of classified objects and the *classification* as applying these rules to new objects of unknown class.

Mathematical description of the previous problem can be formulated in probabilistic terms. Consider finite number $K \geq 2$ of groups of objects. Each object can be represented by $d \in \mathbb{N}$ numerical characteristics. Each group of objects is characterized by the distribution of the numerical characteristics of its members. We denote these distributions $P_1, \ldots, P_K$. It is natural to assume $P_i \neq P_j$ when

$i \neq j$. The distributions are unknown. Consider further K independent random samples $\boldsymbol{X}_{i,1}, \ldots, \boldsymbol{X}_{i,n_i}$, $i = 1, \ldots, K$, from distributions $P_1, \ldots, P_K$. These random samples (known as the training set) provide only available information on the considered distributions. Any vector $\boldsymbol{x} \in \mathbb{R}^d$, representing an object not included in the training set, is considered to be a realization of random vector from one of the distributions $P_1, \ldots, P_K$, but it is unknown from which of them. There is a need to estimate to which group the object belongs. The goal of the discriminant analysis is to find some general rule which allocates arbitrary d-dimensional real vector lying in a support of at least one distribution to one of the considered distributions (groups respectively). The rule (known as classifier) has a form of some measurable function $d : \mathbb{R}^d \to \{1, \ldots, K\}$. The classifier can not be faultless in many cases. A convenient classifier estimates correctly most objects or fits some other desirable claims. Closer explanation is provided in the next paragraph which is devoted to assessment of classifiers.

3.2 Classifier assessment

A natural desirable property of any classifier is small number of misclassified objects. The probability, that some randomly chosen object will be misclassified by a certain classifier $d(\cdot)$, is said to be the *average misclassification rate of the classifier*. The probability is usually expressed in terms of prior and conditional probabilities as

$$\sum_{i=1}^{K} \pi_i \mathrm{P}\left(d(\boldsymbol{X}) \neq i | \boldsymbol{X} \sim P_i\right),$$

where the first term $\pi_i = \mathrm{P}(\boldsymbol{X} \sim P_i)$ is prior probability that X comes from the i-th group, and the second term $\mathrm{P}\left(d(\boldsymbol{X}) \neq i | \boldsymbol{X} \sim P_i\right)$ is conditional probability of wrong classification given that $\boldsymbol{X}$ comes from the i-th group. The word "average" is used to indicate that objects of all classes are considered.

The classifier, which minimizes average misclassification rate is known as *Bayes minimal error rule* (or sometimes it is called optimal Bayes rule). The rule can be expressed as:

$$d(\boldsymbol{x}) = \arg \max_{i=1\ldots,K} \pi_i f_i(\boldsymbol{x}), \tag{3.1}$$

where π_i is prior probability and $f_i(\cdot)$ is a density function of the i-th distribution. The average misclassification rate of this classifier is known as the *optimal Bayes risk*. It is essential to realize, that any classifier can not have lower average misclassification rate than the optimal Bayes risk. The optimal Bayes risk is an attribute of considered distributions and distances between them. While Bayes risk is equal to zero for two distributions with disjoint supports, it is nearly one half for two normal distributions with the same mean and only a small difference in variances or in the case of two normal distributions with the same variances and only small shift in location.

Densities $f_i(\cdot)$ in expression (3.1) are usually unknown and need to be estimated. Sometimes prior probabilities are also unknown. In such a case they are usually estimated by proportions of observations in the training set. The problem of the classifier construction might seem to be reduced to the problem of density estimation. Majority of traditional methods are based indeed on density estimation.

Some of these methods are parametric (considering certain family of distributions, based on parameter estimation), some are nonparametric (these methods weaken the assumptions on considered distributions), some are semiparametric.

Although minimizing average misclassification rate might seem well-founded, it can be spurious in some situations. For example if we consider two distributions with markedly different prior probabilities (lets say $\pi_1 = 0.99$ and $\pi_2 = 0.01$). Then trivial classifier, which classifies all objects into the first group has average misclassification rate equal to 0.01. Although the misclassification rate is low, the classifier is useless. We need another criterion in this situation.

Previously mentioned weakness of minimal average misclassification rate lead to construction of other criteria. Probably the most widely used is a concept of cost function, assigning "costs" of wrong classification of objects. Another strategy (known as the minimax) is based on minimizing maximum possible risk. We restrict our attention on the assessment of classifiers according to their average misclassification rate in this work.

3.3 Simple depth-based methods of discrimination

During the last ten years quite a lot of effort has been put into development of an alternative nonparametric approach, which uses methodology of data depth for solving the discrimination problem. The idea of using data depth for discrimination was firstly introduced in paper by Christmann and Rousseeuw in 2001 ([8]). Number of researchers followed and broaden the idea of the pioneering paper. This section aims to provide an overview of simple discrimination methods based on data depth.

3.3.1 Maximal depth classifier

Maximal depth classifier is probably the most widely used classifier based on data depth. It was used for example in works of Jörnsten [28], Hartikainen and Oja [20] or Mosler and Hoberg [29]. A detailed inspection of the method is provided in a paper by Ghosh and Chaudhuri [18].

The classifier is based on a simple idea of assigning a new observation (represented by vector $\boldsymbol{x}$) to the distribution, with respect to which it has maximal depth. An arbitrary depth function can be used. Of course, different classifiers are gained by using different depth functions.

$$d(\boldsymbol{x}) = \arg \max_{j=1\ldots,K} D(\boldsymbol{x}; P_j). \tag{3.2}$$

Since the theoretical depth is usually unknown, empirical version based on the data from training set is used:

$$d(\boldsymbol{x}) = \arg \max_{j=1\ldots,K} D(\boldsymbol{x}; \widehat{P}_j), \tag{3.3}$$

where $D(\boldsymbol{x}; \widehat{P}_j)$ is a depth of $\boldsymbol{x}$ with respect to empirical distribution of the j-th distribution, which is based on the appropriate points from training set $(\boldsymbol{X}_{j,1}, \ldots, \boldsymbol{X}_{j,n_j})$.

The maximal depth classifier is known to be asymptotically optimal (it has the lowest possible average misclassification rate) in some situations. Gosh and Chaudhuri showed asymptotical optimality of the classifier if the considered distributions are elliptically symmetric with the density function strictly decreasing in every direction from its centre of symmetry, differ only in location (that is they have equal dispersion and are of the same type), and have equal prior probabilities. In addition, the used depth function must satisfy some properties in this settings.

Formally, it is assumed that:

(P1) $f_i(\boldsymbol{x}) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} g\left((\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)$, where $f_i(\cdot)$ is a density function of the distribution P_i (for all $i = 1, \ldots, K$), $\boldsymbol{\mu}_i$ is the mean of P_i and $\boldsymbol{\Sigma} > 0$ is its variance matrix,

(P2) $g(cx) < g(x)$ for arbitrary $x \in \mathbb{R}^+$, for $c > 1$,

(P3) $\pi_1 = \pi_2 = \ldots = \pi_K (= 1/K)$.

Following lemma shows that the maximal depth classifier is equivalent to the classifier which minimizes Mahalanobis distance in the considered case.

Lemma 3.1 *Assume (P1) - (P2). Consider any depth function which have all properties stated in Definition 1.1 in the considered situation. Then for any* $\boldsymbol{x} \in \mathbb{R}^d$

$$\arg\max_{j=1\ldots,K} D(\boldsymbol{x}; P_j) = \arg\min_{j=1\ldots,K} M_j(\boldsymbol{x}),$$

where $M_j(\boldsymbol{x})$ *denotes Mahalanobis distance of* $\boldsymbol{x}$ *from the* P_j*.*

Proof: Affine invariance of the depth function ensures the following equality:

$$\arg\max_{j=1\ldots,K} D(\boldsymbol{x}; P_j) = \arg\max_{j=1\ldots,K} D\left(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}_j); P_0\right),$$

where P_0 is standardized elliptically symmetric distribution of the considered type, that is distribution with the density function $f(\boldsymbol{x}) = g(\boldsymbol{x}^T\boldsymbol{x})$. Using the affine invariance property once more, we get

$$\arg\max_{j=1\ldots,K} D(\boldsymbol{x}; P_j) = \arg\max_{j=1\ldots,K} D\left(\left(\left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right\|, 0, \ldots, 0\right)^T; P_0\right).$$

Now we compare depths of K points lying on the common line with respect to the same distribution P_0. Properties of maximality at centre and monotonicity relative to the deepest point ensures

$$\arg\max_{j=1\ldots,K} D(\boldsymbol{x}; P_j) = \arg\min_{j=1\ldots,K} \left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right\| = \arg\min_{j=1\ldots,K} M_j(\boldsymbol{x}).$$

□

Notice that the claim holds even if equal variance matrices *are not* assumed.

Theorem 3.2 *Assume (P1) - (P3). Consider any depth function which has all properties stated in Definition 1.1 in the considered situation. Then the average misclassification rate of an empirical depth-based classifier (3.3) converges to the optimal Bayes risk as* $\min(n_1, \ldots, n_K) \to \infty$*.*

Proof: The optimal Bayes rule in the considered settings can be simplified:

$$\arg\max_{j=1\ldots,K} \pi_j f_j(\boldsymbol{x}) = \arg\max_{j=1\ldots,K} \frac{1}{K}\boldsymbol{\Sigma}^{-1/2} g\left(M_j^2(\boldsymbol{x})\right) = \arg\min_{j=1\ldots,K} M_j(\boldsymbol{x}).$$

Theoretical maximal depth classifier is thus equivalent to the optimal Bayes classifier (see Lemma 3.1). The claim follows from the uniform convergence of empirical depth functions.

□

Assumptions that are needed for optimality of the maximal depth classifier are very restrictive. Following example shows, that the maximal depth classifier is not optimal when nonequal dispersions are considered:

Example Let us consider two bivariate normal distributions with equal prior probabilities $P_1 = \mathrm{N}\left((0,0)^T, 4\boldsymbol{I}\right)$, and $P_2 = \mathrm{N}\left((1,0)^T, \boldsymbol{I}\right)$, where $\boldsymbol{I}$ denotes 2×2 identity matrix. Denote the new observation $\boldsymbol{x} = (x_1, x_2)^T$.

In this case the optimal Bayes rule has the following form: $d(\boldsymbol{x}) = 2$ iff $(x_1 - 4/3)^2 + x_2^2 < 4/9 + 16/3 \ln 2$. Expected misclassification rate for the group 1 is about 0.3409, for group 2 it is about 0.1406, hence the optimal Bayes risk is about 0.2408.

The theoretical maximal depth classifier, which is equivalent to the classifier minimizing Mahalanobis distance, has the form: $d(\boldsymbol{x}) = 2$ iff $(x_1 - 4/3)^2 + x_2^2 < 4/9$. Expected misclassification rate is 0.0435 for group 1 and 0.8104 for group 2, yielding the average misclassification rate of about 0.4270, which is much higher than the optimal Bayes risk. (The expected misclassification rates were enumerated by the numeric integration of densities).

Similar problems occur when prior probabilities are not equal.

3.3.2 Classifiers for skewed data

Hubert and van der Veeken [25] have been giving special attention to the classification rules for skewed distributions. Their classifier can be considered as a special case of the maximal depth classifier, which uses the modified projection depth function.

The projection depth (described in detail in Section 1.2.5) links up the notions of depth and outlyingness. It was used for classification purposes by Kosiorowski [33] or Dutta and Ghosh [11]. Its construction can be divided into three steps:

1. Consider any measure of outlyingness in one-dimensional case. The outlyingness of a point $x \in \mathbb{R}^1$ with respect to a probability distribution P_X of a random variable X can be defined for example as

$$O(x; P_X) = \frac{x - \mathrm{med}(X)}{\mathrm{MAD}(X)},$$

where med denotes the median and MAD denotes median absolute deviation: $\mathrm{MAD}(X) = \mathrm{med}(|X - \mathrm{med}(X)|)$.

2. In multidimensional case, the projection pursuit technique can be applied. The outlyingness of a point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to a probability distribution $P_{\boldsymbol{X}}$ of a random vector $\boldsymbol{X}$ is defined as

$$O(\boldsymbol{x}; P_{\boldsymbol{X}}) = \sup_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} O(\boldsymbol{u}^T\boldsymbol{x}; P_{\boldsymbol{u}^T\boldsymbol{X}}).$$

3. The projection depth is defined as

$$D(\boldsymbol{x}; P_{\boldsymbol{X}}) = \frac{1}{1 + O(\boldsymbol{x}; P_{\boldsymbol{X}})}.$$

This construction can be used for an arbitrary notion of one-dimensional outlyingness. The outlyingness of a point describes how far the point lies from the centre of the data. Usually it does not matter whether the data point is smaller or larger than the centre point. This property might be considered undesirable when the distribution is skewed. Hubert and van der Veeken [24] proposed the outlyingness measure (called adjusted outlyingness), which takes into account the skewness of a distribution. They used the notion of medcouple, a robust alternative to the classical skewness coefficient, proposed by Brys et al. [6]:

Definition 3.1 *Let X be a univariate random variable, which has continuous distribution P_X with the unique median* $\mathrm{med}(X)$. *The medcouple of the distribution P_X is defined as*

$$MC(P_X) := \mathrm{med}\left(\frac{(X_U - \mathrm{med}(X)) - (\mathrm{med}(X) - X_L)}{X_U - X_L}\right),$$

where X_U and X_L are independent variables which come from truncated P distributions: X_U has conditional distribution of X given $X >$ $\mathrm{med}(X)$ *and X_L has conditional distribution of X given $X <$* $\mathrm{med}(X)$.

Obviously $MC(P_X) = 0$ for symmetric distributions, $MC(P_X) > 0$ for right skewed distributions and $MC(P_X) < 0$ for left skewed distributions. Brys et al. have shown some basic properties like location and scale invariance of the medcouple: $MC(P_{aX+b}) = MC(P_X)$ for any $a > 0$ and $b \in \mathbb{R}$. Inverting distribution causes inverting of medcouple: $MC(P_{-X}) = -MC(P_X)$. A procedure `mc` estimating the medcouple exists in R-package `robustbase`.

The adjusted outlyingness is defined as follows:

Definition 3.2 *Let X be a one-dimensional random variable, which has continuous distribution P_X with the median* $\mathrm{med}(X)$, *lower quartile $Q_1(X)$, upper quartile $Q_3(X)$, interquartile range $IQR(X) = Q_3(X) - Q_1(X)$ and medcouple $MC(P_X) \geq 0$. The adjusted outlyingness of a point $x \in \mathbb{R}^1$ with respect to the distribution of random variable X is defined as*

$$AO(x; P_X) = \begin{cases} \frac{x - \mathrm{med}(X)}{c_U - \mathrm{med}(X)} & \text{if } x > \mathrm{med}(X) \\ \\ \frac{\mathrm{med}(X) - x}{\mathrm{med}(X) - c_L} & \text{if } x < \mathrm{med}(X), \end{cases} \tag{3.4}$$

where
$c_L = Q_1(X) - 1.5\mathrm{e}^{-4MC(P_X)}IQR(X)$ *and* $c_U = Q_3(X) + 1.5\mathrm{e}^{3MC(P_X)}IQR(X)$.

Note, that the adjusted outlyingness is defined for right skewed distributions (for which $MC(P_X) > 0$). If the distribution is left skewed, we compute the adjusted outlyingness of $-x$ with respect to the distribution of $-X$. As far as we know, constants -4 and 3 in the definition of c_L and c_U were proposed ad hoc.

Example We illustrate the definition by an example. Consider a random variable X with lognormal distribution with parameters $\mu = 0, \sigma = 2/3$. Then $\text{med}(X) = 1$, quartiles are tabulated ($Q_1 = 0.64, Q_3 = 1.57$), the medcouple is about 0.28. The positive value of the medcouple indicates right skewed distribution. When we consider two points equally distant from the median, the adjusted outlyingness of the point smaller than median is higher than the adjusted outlyingness of the point on the right side of the median. For example, $AO(0.25; P_X) = 0.92$, while $AO(1.75; P_X) = 0.20$. The example is illustrated in Figure 3.1.

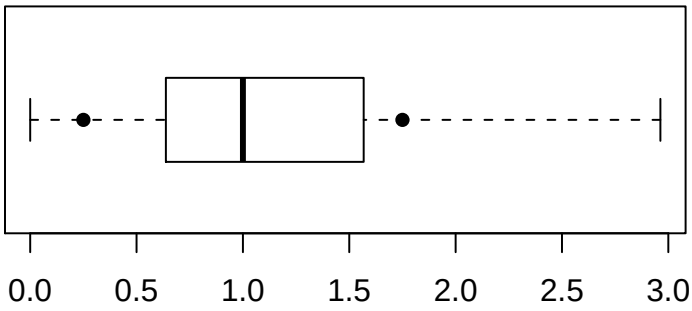


Figure 3.1: Boxplot of $X \sim logN(0, 2/3)$ with two points equally distant from median: 0.25 and 1.75.

The proposed classifier consists in minimization of the adjusted outlyingness (or maximization of the corresponding projection depth). It can be considered to be a special case of maximal depth classifier (3.2). It works very well in situations presented by Hubert and Van der Veeken ([25]). They consider two distributions which come from the family of skewed normal distributions. The family of skewed normal distributions, introduced by Azzalini (see [1]), includes normal distributions and some others, that are derived from normal distribution by its "skewing". The formal definition follows:

Definition 3.3 *A d-dimensional random vector $\boldsymbol{X}$ has a central-skewed-normal distribution with variance matrix $\boldsymbol{\Sigma}_0$ and skewness-regulating parameter $\boldsymbol{\alpha} \in \mathbb{R}^d$, if its density function is of form*

$$f(\boldsymbol{x}) = 2\phi_d(\boldsymbol{x}, \boldsymbol{\Sigma}_0)\Phi(\boldsymbol{\alpha}^T\boldsymbol{x}),$$

where $\phi_p(\cdot, \boldsymbol{\Sigma}_0)$ is a density function of d-dimensional normal distribution with zero mean and variance matrix $\boldsymbol{\Sigma}_0$; $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. We write $\boldsymbol{X} \sim SN_d(\boldsymbol{0}, \boldsymbol{\Sigma}_0, \boldsymbol{\alpha})$.

A d-dimensional random vector $\boldsymbol{Y}$ has a skewed-normal distribution with mean $\boldsymbol{\mu}$, variance matrix $\boldsymbol{\Sigma}$ and skewness-regulating parameter $\boldsymbol{\alpha} \in \mathbb{R}^d$ if

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\omega}^T\boldsymbol{X},$$

where $\boldsymbol{X} \sim SN_d(\boldsymbol{0}, \boldsymbol{\Sigma}_0, \boldsymbol{\alpha})$, and $\boldsymbol{\Sigma} = \boldsymbol{\omega}^T\boldsymbol{\Sigma}_0\boldsymbol{\omega}$. We write $\boldsymbol{Y} \sim SN_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$.

Example Consider now discrimination problem with two bivariate distributions from the family of skewed-normal distributions:

- Distribution $P_1 = N_2(\mathbf{0}, \boldsymbol{I})$ has the density function $f_1(\boldsymbol{x}) = \phi(x_1)\phi(x_2)$.
- Distribution $P_2 = SN_2(-2 \cdot \mathbf{1}, \boldsymbol{I}, 5 \cdot \mathbf{1})$ has the density function $f_2(x) = 2\phi(x_1 + 2)\phi(x_2 + 2)\Phi\left(5[(x_1 + 2) + (x_2 + 2)]\right)$.

We suppose equal prior probabilities.

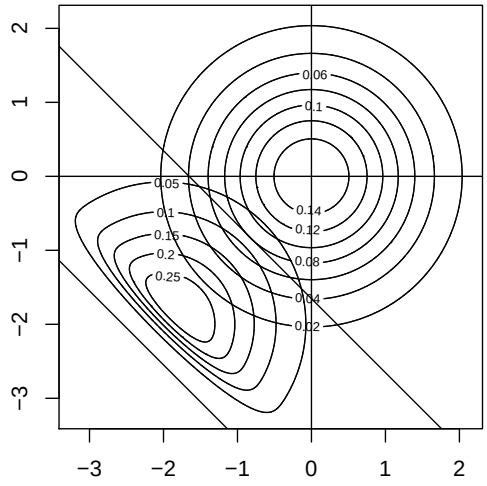


Figure 3.2: Levelsets of density of normal $N_2(\mathbf{0}, \boldsymbol{I})$ and skewed-normal $SN_2(-2 \cdot \mathbf{1}, \boldsymbol{I}, 5 \cdot \mathbf{1})$ distribution; Bayes discriminant rule is depicted by diagonal lines.

The Bayes rule allocates $\boldsymbol{x} = (x_1, x_2)^T$ to:

group 1 if $x_2 > -x_1 - 1.65$ or $x_2 < -x_1 - 4.55$,

group 2 if $x_2 < -x_1 - 1.65$ and $x_2 > -x_1 - 4.55$.

Where the constants -1.65 and -4.55 are approximative values.

Misclassification rate for the group 1 can be expressed as $\Phi(q_2/\sqrt{2}) - \Phi(q_1/\sqrt{2})$. It is about 0.1205. Misclassification rate for the group 2 can be find out by simulation. It is about 0.097. Considering equal priors the average misclassification rate is approximately 0.109.

The average misclassification rate for the maximal depth classifier which uses adjusted outlyingness is near the optimal Bayes risk. We estimated the rate to be 0.111. The example is described in detail in Section 3.7.3.

The low average misclassification rate in previous example, similarly as other results presented by Hubert and Van der Veeken, are very impressive. Nevertheless, problems will arise when considering two distributions with more evident difference in dispersion.

Example Recall the example presented in Section 3.3.1. We consider two bivariate normal distributions with equal prior probabilities $P_1 = \mathrm{N}\left((0,0)^T, 4\boldsymbol{I}\right)$, and $P_2 = \mathrm{N}\left((1,0)^T, \boldsymbol{I}\right)$. In this case both distributions are symmetric and hence $MC = 0$.

Adjusted outlyingness of any one-dimensional projection is thus reduced to the form

$$AO(\boldsymbol{u}^T\boldsymbol{x}; P_{\boldsymbol{u}^T\boldsymbol{X}}) = \frac{\boldsymbol{u}^T\boldsymbol{x} - \mathrm{med}(\boldsymbol{u}^T\boldsymbol{x})}{2IQR(\boldsymbol{u}^T\boldsymbol{x})}.$$

The maximal depth classifier based on adjusted outlyingness has the form: $d(\boldsymbol{x}) = 2$ iff $(x_1 - 4/3)^2 + x_2^2 < 4/9$, which was shown to be far away from optimality in previous section.

The result is not surprising, because the projection depth (or more precisely the depth base on adjusted outlyingness), is affine invariant notion of depth and hence general results described in Section 3.3.1 hold.

3.3.3 Maximal central area classifier

A new idea how to use the notion of data depth for solving discrimination problem was proposed by Billor et al. in 2008 (see [3]). They realized that the range of possible depth values can vary among competing populations. For example, the deepest point of a symmetric distribution has halfspace depth equal to 0.5, but the deepest point of an asymmetric distribution has smaller depth. This simple consideration lead to the idea of classifying a new object $\boldsymbol{x}$ into the group, which has the highest expected proportion of points with smaller depth than the depth of $\boldsymbol{x}$:

$$d(\boldsymbol{x}) = \arg\max_{i=1,\dots,K} \mathrm{P}\left(D(\boldsymbol{X}_i; P_i) \leq D(\boldsymbol{x}; P_i)\right), \tag{3.5}$$

where $D(\boldsymbol{x}; P_i)$ is the depth of point $\boldsymbol{x}$ with respect to the i-th distribution, and $D(\boldsymbol{X}_i; P_i)$ is random variable, which can be described as the depth of point $\boldsymbol{X}_i$ randomly sampled from P_i with respect to the distribution P_i.

In other words, the classifier chooses such a distribution P_i, for which the value $D(\boldsymbol{x}; P_i)$ is the highest quantile of $D(\boldsymbol{X}_i; P_i)$. It can be also expressed in terms of central areas: denoting $C_{q(i)}$ the smallest central area of the i-th distribution which includes $\boldsymbol{x}$ (as in Definition 1.2), the classifier allocates $\boldsymbol{x}$ to the distribution P_i for which $q(i)$ is maximal. The classifiers could be also expressed in terms of ranks, as they were defined in Section 1.3. Following formula is equivalent to (3.5):

$$d(\boldsymbol{x}) = \arg\max_{i=1,\dots,K} r_{P_i}(\boldsymbol{x}).$$

The theoretical probability in expression (3.5) is practically always unknown and must be estimated. The estimation based on training set leads to the simple rule:

$$d(\boldsymbol{x}) = \arg\max_{i=1,\dots,K} \frac{1}{n_i} \sum_{j=1}^{n_i} I\left(D(\boldsymbol{X}_{i,j}, \widehat{P}_i) \leq D(\boldsymbol{x}; \widehat{P}_i)\right), \tag{3.6}$$

where $\boldsymbol{X}_{i,j}$ is the j-th point from the i-th group of training set, n_i denotes number of points from the i-th group in training set, $\widehat{P}_i$ denotes empirical version of P_i based on the data from training set and $I(\cdot)$ is an indicator function equal to one if the condition in argument is satisfied and is zero otherwise.

The classifier was originally derived by Billor et al. from the idea of transvariations and it was called the depth transvariation classifier. However, the derivation was somewhat confusing. It was based on the following definition of the transvariation probability:

Definition 3.4 *The transvariation probability between distribution P of a univariate random variable Y and a constant $c \in \mathbb{R}$ is defined as*

$$\tau(c) = \mathrm{P}\left\{(Y - c)\,(\mu - c) \leq 0\right\}, \tag{3.7}$$

where μ is some location parameter of the distribution P.

It is not clear which location parameter should be used.

We consider random variables $Y_i = D(\boldsymbol{X}_i; P_i)$ and a constants $c_i = D(\boldsymbol{x}; P_i) \in \mathbb{R}$; μ_i are some location parameters of the distributions of Y_i. The idea used by Billor et al. is to allocate a new observation $\boldsymbol{x}$ to the group, for which the transvariation probability between the constant $D(\boldsymbol{x}; P_i)$ and the distribution of $D(\boldsymbol{X}_i, P_i), \boldsymbol{X}_i \sim P_i$ is maximal. Billor et al. considered μ_i to be the depth of the deepest point with respect to P_i. In this case $\mu_i - c_i \geq 0$ and the formula (3.7) simplifies to $\tau(c_i) = \mathrm{P}\left(Y_i - c_i \leq 0\right)$, which leads to the classifier (3.5). Using other location parameter like mean or median of the $D(\boldsymbol{X}_i, P_i)$ would lead to poor classifier, because points with the high depth would have low transvariation probability.

We propose more straightforward justification of the classifier (3.5). Natural location parameter of a symmetric one-dimensional distribution P is its point of symmetry. The transvariation probability between distribution P and a point $c \in \mathbb{R}$ is increasing with the decreasing distance of c from the centre of symmetry μ. The same idea can be applied in multidimensional space $\mathbb{R}^d, d > 1$. Such an extension, which uses depth of the point as a measure of distance from the centre of symmetry, leads directly to the classifier (3.5).

Theoretical properties of the classifier (3.5) are not studied in [3]. However, we can derive some results quite easily under similar assumptions as in Section 3.3.1 (elliptical symmetry, difference in location and equal priors):

Theorem 3.3 *Assume (P1) - (P3) from Section 3.3.1. Consider any depth function which has all properties stated in Definition 1.1 in the considered situation. Then the classifier (3.5) is equivalent to the Bayes rule (3.1).*

Proof: It was already shown in the proof of Theorem 3.2 that the optimal Bayes rule in the considered settings has form:

$$d(\boldsymbol{x}) = \arg\min_{i=1\ldots,K} M_i(\boldsymbol{x}),$$

where $M_i(\boldsymbol{x})$ denotes Mahalanobis distance of $\boldsymbol{x}$ from P_i. By repeating the steps in proof of Lemma 3.1 we can show that the depth $D(\boldsymbol{x}; P_i)$ is a decreasing function of the Mahalanobis distance $M_i(\boldsymbol{x})$ in the considered case. Hence

$$\mathrm{P}\left(D(\boldsymbol{X}_i; P_i) \leq D(\boldsymbol{x}; P_i)\right) = \mathrm{P}\left(M_i(\boldsymbol{X}_i) \geq M_i(\boldsymbol{x})\right), \tag{3.8}$$

where $M_i(\boldsymbol{X}_i)$ is a random variable, which is the Mahalanobis distance of random variable $\boldsymbol{X}_i$ from the distribution P_i. Now it suffices to realize that the distribution of $M_i(\boldsymbol{X}_i)$ is the same for all i (it does not depend on i). This is clear from the following equation:

$$M_i(\boldsymbol{X}_i)^2 = (\boldsymbol{X}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_i) = \boldsymbol{Y}_i^T \boldsymbol{Y}_i,$$

where $\boldsymbol{Y}_i = \boldsymbol{\Sigma}_i^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu}_i)$ has the same distribution for all $i = 1, \ldots, K$: elliptically symmetric with zero mean and identity variance matrix. Now it is clear from (3.8) that

$$\arg \max_{i=1,\ldots,K} \mathrm{P}\left(D(\boldsymbol{X}_i; P_i) \leq D(\boldsymbol{x}; P_i)\right) = \arg \min_{i=1,\ldots,K} M_i(\boldsymbol{x}).$$

□

The maximal central classifier (3.5) is not optimal in the case of unequal dispersions, because it remains equivalent to the classifier $d(\boldsymbol{x}) = \arg\min_{i=1,\ldots,K} M_i(\boldsymbol{x})$, maximal depth classifier (3.2) respectively, but these classifiers are not equivalent to the Bayes optimal rule any more, as was shown in Section 3.3.1.

The improvement of maximal central classifier showed by Billor et al. is due to the use of L_1–depth, which is not affine invariant. The difference between classifiers (3.3) and (3.6) can be visible when the considered distributions are of different type. Then the later classifier comprehends possibly different ranges of depth functions.

3.3.4 Problem of different dispersions

Both maximal depth classifier and maximal central classifier are not optimal when the considered distributions differ in dispersion. In previous sections distributions $P_1, \ldots, P_K$ are assumed to be of the same type and differ only in location parameter. The class of problems, that can be satisfactorily solved by use of classifier (3.3) or (3.6), is thus quite narrow.

The problems of classifiers based on data depth arise from the discrepancy between the depth and the density function. The classifiers considered in this section are based on data depth; the optimal Bayes classifier is based on density function. While the depth function is affine invariant, the density function does not have this property. Recall how the density function and the (affine invariant) depth function change after an affine transformation: Consider a d-dimensional random vector $\boldsymbol{X}$ with a density function f. Consider a regular $d \times d$ matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b} \in \mathbb{R}^d$. The density function of a random vector $Y := \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ is denoted by g. Then it holds:

$$\begin{aligned} g(\boldsymbol{y}) &= f(\boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b}))\,|\boldsymbol{A}|^{-1} \\ D(\boldsymbol{y}; \boldsymbol{Y}) &= D(\boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b}); \boldsymbol{X}). \end{aligned}$$

When $|\boldsymbol{A}| \neq 1$, the transformation "changes" the density, but "does not change" the depth. This discrepancy may lead to serious problems of depth-based classifiers. Following simple example illustrates problems arising from affine transformation such as a change of scale:

Example Consider a normal distribution $P_1 = N(0, \sigma_1^2)$. When the scale is changed, distribution P_1 is transformed to $P_2 = N(0, \sigma_2^2)$. Assume $0 < \sigma_1 < \sigma_2$. Equal prior probabilities of P_1 and P_2 can be considered. For any $x \neq 0$ it holds $D(x; P_1) < D(x; P_2)$. Both (3.2) and (3.5) classify all new objects to P_2 with probability one. The average misclassification rate is thus equal to $1/2$. The rule is as bad as a random assigning to the distributions (coin tossing). The Bayes rule

assigns a new observation x to P_1 iff $|x| < x_0 = \sigma_1\sigma_2 \left[\frac{\ln\sigma_2^2 - \ln\sigma_1^2}{\sigma_2^2 - \sigma_1^2}\right]^{1/2}$. The average misclassification rate is $1/2 - [\Phi(-x_0/\sigma_2) - \Phi(-x_0/\sigma_1)] < 1/2$. For example, if $\sigma_1 = 1$ and $\sigma_2 = 4$, the Bayes optimal risk is about 0.21.

3.4 Advanced depth-based methods of discrimination

In this section we describe two classifiers based on depth, which overcome the problems with different dispersions. They both are based on the assumption that there exists a relationship between the depth and the density function. The first one uses one-dimensional kernel density estimation. This approach was proposed by Gosh and Chaudhuri [18]. The later approach modifies the k-nearest-neighbour method. This modification has not been published so far.

3.4.1 Dealing with different dispersions

The clue to construction of an effective classifier based on data depth lies in the existence of a relationship between the depth and the density function. Formally, it is assumed that the density function $f_i(\cdot)$ corresponding to the distribution P_i can be expressed as a function of a depth $D(\cdot; P_i)$ for all $i = 1, \ldots, K$:

$$f_i(\boldsymbol{x}) = h_i(D(\boldsymbol{x}; P_i)), \quad i = 1, \ldots, K. \tag{3.9}$$

In such a case the depth function can be used to estimate the density function. In this way an empirical version of the Bayes rule can be constructed. The Bayes classifier (3.1) in this case has a form:

$$d(\boldsymbol{x}) = \arg\max_{i=1\ldots,K} \pi_i h_i(D(\boldsymbol{x}; P_i)). \tag{3.10}$$

Note that the functions $h_i(\cdot)$ need not to be the same for all distributions. Recall the example from Section 3.3.4. We consider $P_1 = N(0, \sigma_1^2)$ and $P_2 = N(0, \sigma_2^2)$. When considering the halfspace depth, the density function $f_1(\cdot)$ corresponding to the P_1 and the density function $f_2(\cdot)$ corresponding to P_2 can be expressed as

$$\begin{aligned} f_1(x) &= \phi(x/\sigma_1)/\sigma_1 = \phi(-x/\sigma_1)/\sigma_1 = \phi\left(\Phi^{-1}\left(D(x; P_1)\right)\right)/\sigma_1, \\ f_2(x) &= \ldots = \phi\left(\Phi^{-1}\left(D(x; P_2)\right)\right)/\sigma_2, \end{aligned}$$

where $\phi(\cdot)$ denotes the density function of the standardized normal distribution and $\Phi(\cdot)$ denotes the cumulative distribution function of the standardized normal distribution. Hence we have $h_1(\cdot) = \phi\left(\Phi^{-1}(\cdot)\right)/\sigma_1$, while $h_2(\cdot) = \phi\left(\Phi^{-1}(\cdot)\right)/\sigma_2$.

On the other hand, $h_i(\cdot)$ might be independent on i under a certain circumstances, for example if we consider elliptically symmetric distributions of the same type, which differ only in location (have equal dispersions).

Two general concepts are widely used in density estimation for the purposes of discrimination. The kernel density estimation and the k-nearest-neighbour

method. They both can take advantage of the idea of the data depth. The methods are based on the following approximation:

$$\mathrm{P}(\boldsymbol{X} \in L(\boldsymbol{x})) = \int_{L(\boldsymbol{x})} f(\boldsymbol{y}) d\boldsymbol{y} \cong f(\boldsymbol{x}) \cdot \lambda_d \left(L(\boldsymbol{x}) \right), \tag{3.11}$$

where $L(\boldsymbol{x})$ is some neighbourhood of $\boldsymbol{x} \in \mathbb{R}^d$, $f(\cdot)$ is the density function of a random vector $\boldsymbol{X}$ and $\lambda_d \left(L(\boldsymbol{x}) \right)$ is a d-dimensional Lebesgue measure of $L(\boldsymbol{x})$ - the volume of $L(\boldsymbol{x})$. The density can be estimated by estimation of the following ratio:

$$f(\boldsymbol{x}) \cong \mathrm{P}(\boldsymbol{X} \in L(\boldsymbol{x})) / \lambda_d \left(L(\boldsymbol{x}) \right). \tag{3.12}$$

The approximation is appropriate as far as the density function is nearly constant (equal approximately to $f(\boldsymbol{x})$) at the neighbourhood $L(\boldsymbol{x})$. Classical methods work with a neighbourhood defined as

$$L(\boldsymbol{x}) = \left\{ \boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{y}\| < \eta \right\}, \tag{3.13}$$

for some small positive constant $\eta \in \mathbb{R}$. The definition is simple, but it leads to the problems connected with sparsity of data in high dimensional space (known as the curse of dimensionality). To prevent these problems, the relationship (3.9) can be used. Since levelsets of depth correspond to levelsets of density, $f(\cdot)$ is constant whenever $D(\cdot)$ is constant. We can define the "neighbourhood" of $\boldsymbol{x}$, as the set of all points with the depth close to the depth of $\boldsymbol{x}$. Since the depth is always related to the distribution, we talk about "distributional neighbourhood". Formally, we define

$$L(\boldsymbol{x}; P) = \left\{ \boldsymbol{y} \in \mathbb{R}^d : |D(\boldsymbol{x}; P) - D(\boldsymbol{y}; P)| < \eta \right\}, \tag{3.14}$$

for some small positive constant $\eta \in \mathbb{R}$. In other words $\boldsymbol{y} \in L(\boldsymbol{x}; P)$ iff $D(\boldsymbol{y}; P) \in U_\eta(D(\boldsymbol{x}; P))$, where $U_\eta(D(\boldsymbol{x}; P))$ is classical one-dimensional neighbourhood of $D(\boldsymbol{x}; P)$.

Notice that the definitions (3.13) and (3.14) are not equivalent. The later definition constitutes an alternative approach to the notion of a neighbourhood. The difference between the classical neighbourhood and the distributional neighbourhood of a point is illustrated in Figure 3.3. In the considered example, levelsets of the bivariate normal distribution $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix} \right)$ are plotted. The classical and the distributional neighbourhood of a point $\boldsymbol{x} = [2.5, 0.5]$ are plotted.

An advantage of the later approach is that more points, in which the density is similar to the considered point, can be included in the neighbourhood.

3.4.2 A method based on kernel density estimation

Construction of empirical Bayes classifier is based on the estimation of the densities $f_i(\boldsymbol{x})$, $i = 1, \ldots, K$. The estimates are usually based on the approximation (3.12). Denote $L_i(\boldsymbol{x}) = \left\{ \boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{y}\| < \eta_i \right\}, i = 1, \ldots, K$, the classical neighbourhoods of $\boldsymbol{x}$ used for the estimation of $f_i(\boldsymbol{x})$.

The first usual way how to proceed from (3.12), is to put $\eta_1 = \ldots = \eta_K =: \eta$. Consequently $L_1(\boldsymbol{x}) = \ldots = L_K(\boldsymbol{x}) =: L(\boldsymbol{x}; \eta)$ and maximization of $\widehat{\pi}_i \widehat{f}_i(\boldsymbol{x})$ is

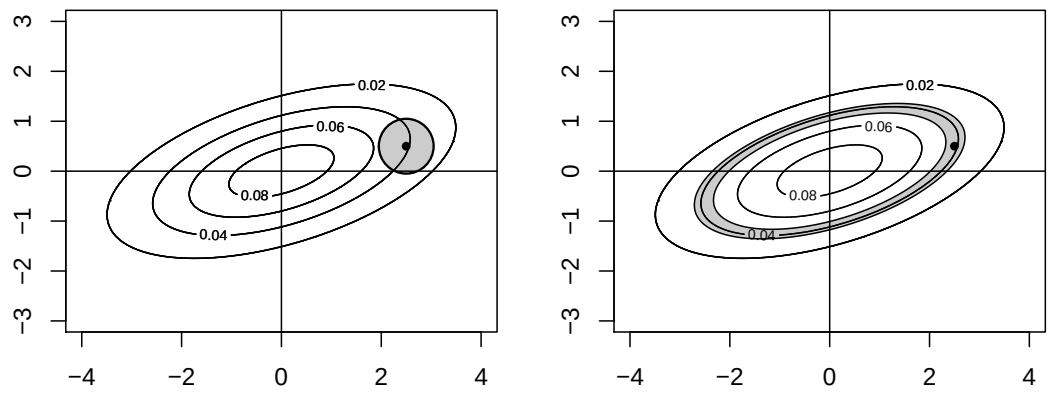


Figure 3.3: A classical neighbourhood of a point $\boldsymbol{x} = [2.5, 0.5]$ (left) and a distributional neighbourhood of the same point (right), when the considered distribution is centralized bivariate normal.

thus equivalent to the maximization of number of points from particular groups of the training set included in $L(\boldsymbol{x}; \eta)$:

$$d(\boldsymbol{x}) = \arg \max_{i=1\ldots,K} \# \left\{ j : \boldsymbol{X}_{i,j} \in L(\boldsymbol{x}; \eta) \right\}.$$

Considering the function $Ker : \mathbb{R}^d \to \mathbb{R}^1$

$$Ker(\boldsymbol{z}) = \begin{cases} 0 & \text{if } \|\boldsymbol{z}\| > 1 \\ 1 & \text{if } \|\boldsymbol{z}\| \leq 1, \end{cases}$$

one can estimate the density functions in $\boldsymbol{x}$ by

$$\widehat{f}_i(\boldsymbol{x}) = \frac{c}{\eta^d} \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} Ker\left(\frac{\boldsymbol{x} - \boldsymbol{X}_{i,j}}{\eta} \right), \ i = 1, \ldots, K,$$

where c is a constant independent on i. The method is usually generalized by allowing $Ker(\cdot)$ to be any appropriate kernel function.

This approach is quite common in one-dimensional case. However, in higher dimensions it has serious problems with sparsity of data. Large η must be chosen to ensure enough points lying inside the neighbourhood $L(\boldsymbol{x}; \eta)$. Large values of η violate the assumption of nearly constant density-level on the neighbourhood $L(\boldsymbol{x}; \eta)$.

An improvement can be gained by using distributional neighbourhood instead of the classical neighbourhood: $L(\boldsymbol{x}, P_i) = \left\{ \boldsymbol{y} \in \mathbb{R}^d : \|D(\boldsymbol{x}; P_i) - D(\boldsymbol{y}; P_i)\| < \eta_i \right\}$. The first step is same as in the traditional approach: we put $\eta_1 = \ldots = \eta_K =: \eta$. A very important difference from the traditional approach is that this equality does not imply equality $L(\boldsymbol{x}, P_1) = \ldots = L(\boldsymbol{x}, P_K)$. By contrast with the traditional approach, K generally different neighbourhoods with different volumes are considered in this case. This fact is illustrated in Figure 3.4, where two different distributional neighbourhoods of the point $\boldsymbol{x} = [2.5, 0.5]$ are plotted. These neighbourhoods are based on the halfspace depth of $\boldsymbol{x}$ with respect to the distribution $P_1 = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix} \right)$, $P_2 = N\left(\begin{pmatrix} 4 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$ respectively.

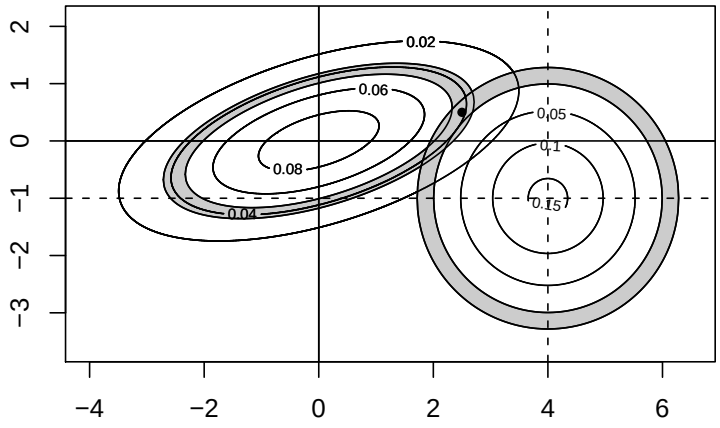


Figure 3.4: Two different distributional neighbourhoods of the point $\boldsymbol{x} = [2.5, 0.5]$.

Different neighbourhoods bring a problem of unequal volumes of neighbourhoods. Two possible solutions have been presented in literature. The first approach (presented by Gosh and Chaudhuri [18]) is based on the assumption that $P_1, \ldots, P_K$ are elliptically symmetric distributions. In that case, one can use a known relationship between volumes of neighbourhoods and the halfspace depth. The other approach (presented by Fraiman et al. [15]) is more general. It is based on the estimate of relationship between the depth and the volumes of the central areas. We discuss both approaches in next paragraphs.

1. Ghosh and Chaudhuri assumed $P_1, \ldots, P_K$ be elliptically symmetric with parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, $i = 1, \ldots, K$. For elliptically symmetric distributions, it is convenient to work with Mahalanobis distance of the point $\boldsymbol{x}$ from the distribution P_i: $M_i(\boldsymbol{x}) = \left[(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right]^{1/2}$, which determines the density $f_i(\boldsymbol{x})$. Let us denote the Radon Nikodym derivative of $\lambda_d \circ M_i^{-1}$ with respect to λ_1 by $\mathcal{L}_i^{(M)}$. Denote the density function of $M_i(\boldsymbol{X})$, where $\boldsymbol{X} \sim P_i$, by $\rho_i^{(M)}(\cdot)$. Since the density of elliptically symmetric distribution is a function of the Mahalanobis distance, we have

$$f_i(\boldsymbol{x}) = h_i^*(M_i(\boldsymbol{x})) = \frac{\rho_i^{(M)}(M_i(\boldsymbol{x}))}{\mathcal{L}_i^{(M)}(M_i(\boldsymbol{x}))}. \tag{3.15}$$

   The form of $\mathcal{L}_i^{(M)}(\cdot)$ is known under the assumption of elliptical symmetry:

$$\mathcal{L}_i^{(M)}(m) = |\boldsymbol{\Sigma}_i|^{1/2} \frac{\pi^{d/2}}{\Gamma(d/2)/2} m^{d-1}.$$

   The formula follows directly from the expression of the volume of d-dimensional ball and the density transformation theorem. The Bayes classifier (3.1) could be expressed in the following form:

$$d(\boldsymbol{x}) = \arg \max_{i=1,\ldots,K} c_i \rho_i^{(M)}(M_i(\boldsymbol{x}))/M_i(\boldsymbol{x})^{d-1}, \tag{3.16}$$

   where c_i are constants, in which π_i and $|\boldsymbol{\Sigma}_i|$ terms are hidden. The approach of Gosh and Chaudhuri described in [18] can be shortly summarized as follows:

- For each group of the training set, we estimate the Mahalanobis distances of all its points from its centre of symmetry.
- The density functions $\rho_i^{(M)}(\cdot)$ can be estimated by one-dimensional kernel density estimation, where the observations are Mahalanobis distances of the points included in the i-th group of the training set estimated in the first step.
- Constant c_1 is defined to be equal to one and constants $c_2, \ldots, c_K$ are estimated by cross-validation, where the average misclassification rate of the rule is minimized when classifying the points from the training set.
- The new observation $\boldsymbol{x}$ is classified according to the empirical version of the rule (3.16), where estimated constants $\widehat{c}_i$, estimated Mahalanobis depth $\widehat{M}_i(\boldsymbol{x})$ and estimated density $\widehat{\rho}_i^{(M)}(\widehat{M}_i(\boldsymbol{x}))$ are used.

Technical details are omitted here. However, we should note that the halfspace depth is used for the robust estimation of the Mahalanobis distance. The estimate is based on the simple relationship between the halfspace depth and the Mahalanobis distance in the case of elliptically symmetric distributions:

Lemma 3.4 *If $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{U}$, where $\boldsymbol{U}$ has a spherically symmetric distribution, then $D(\boldsymbol{x}; \boldsymbol{X}) = 1 - F\left([(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]^{1/2}\right)$, where F is the cumulative distribution function of $\boldsymbol{l}^T\boldsymbol{U}$ for every $\boldsymbol{l}$ with $\|\boldsymbol{l}\| = 1$.*

Proof of the lemma can be found in [18].

2. More general approach proposed by Fraiman et al. [15] does not rely on the assumption of ellipticity. Let $D(\cdot;\cdot)$ be the halfspace depth function; for short denote $D_i := D(\cdot, P_i)$. Assume existence of the Radon Nikodym derivatives of $\lambda_d \circ D_i^{-1}$ with respect to λ_1 and denote them by $\mathcal{L}_i^{(D)}, i = 1, \ldots, K$. Denote the density function of $D_i(\boldsymbol{X})$, where $\boldsymbol{X} \sim P_i$, by $\rho_i^{(D)}(\cdot)$. Now we have

$$f_i(\boldsymbol{x}) = h_i(D(\boldsymbol{x}; P_i)) = \frac{\rho_i^{(D)}(D(\boldsymbol{x}; P_i))}{\mathcal{L}_i^{(D)}(D(\boldsymbol{x}; P_i))}.$$

The form of $\mathcal{L}_i^{(D)}(\cdot)$ is generally not known and need to be estimated. Fraiman et al. proposed to estimate the density function in the following way:

$$\widehat{f}_i(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n_i} \frac{Ker_\eta(D(\boldsymbol{x}; \widehat{P}_i) - D(\boldsymbol{X}_{i,j}; \widehat{P}_i))}{\int Ker_{\eta_f}(D(\boldsymbol{x}; \widehat{P}_i) - D(\boldsymbol{t}; \widehat{P}_i))d\boldsymbol{t}},$$

where $Ker_\eta(\boldsymbol{x}) = Ker(\boldsymbol{x}/\eta)/\eta$ for some kernel Ker and some bandwidths η and η_f. Technical details are omitted here.

3.4.3 Modified k-nearest-neighbour method

In this section we introduce a modification of the well-known k-nearest-neighbour (k-NN) method. Recall that the classical k-NN method is based on the local approximation (3.12), similarly as the classifiers presented in the previous section:

$$\pi_i f_i(\boldsymbol{x}) \cong \pi_i \frac{P(\boldsymbol{X}_i \in L_i(\boldsymbol{x}))}{\lambda_d\left(L_i(\boldsymbol{x})\right)}, \text{where } \ \boldsymbol{X}_i \sim P_i.$$

The probability $P(\boldsymbol{X}_i \in L_i(\boldsymbol{x}))$ can be easily estimated by the proportion of points from the i-th group of the training set lying in $L_i(\boldsymbol{x})$. The prior probabilities π_i can be estimated by relative frequencies of points in particular groups of the training set. Hence the empirical Bayes classifier can be constructed in the following way:

$$\widehat{\pi}_i \widehat{f}_i(\boldsymbol{x}) = \frac{n_i}{n} \frac{k_i}{n_i} \frac{1}{\widehat{\lambda}_d(L_i(\boldsymbol{x}))} = \frac{k_i}{n\widehat{\lambda}_d(L_i(\boldsymbol{x}))}, \tag{3.17}$$

where n_i is the number of points in the i-th group of the training set, n is the total number of points in the training set ($n = n_1 + \ldots, n_K$) and k_i is the number of points from the i-th group of the training set lying in $L_i(\boldsymbol{x})$.

Fix and Hodges [14] proposed to find such a neighbourhood $L(\boldsymbol{x}) = L_1(\boldsymbol{x}) = \ldots = L_K(\boldsymbol{x})$ so that it contains some fixed number of points from the training set. The number of points in the neighbourhood is traditionally denoted by k and therefore the method is called k-nearest-neighbour.

Notice that classical neighbourhoods $L_1(\boldsymbol{x}), \ldots, L_K(\boldsymbol{x})$ can be replaced by distributional neighbourhoods $L(\boldsymbol{x}, P_1), \ldots, L(\boldsymbol{x}, P_K)$. Inspired by the idea of k-NN, we propose to find such distributional neighbourhoods $L(\boldsymbol{x}, P_1), \ldots, L(\boldsymbol{x}, P_K)$ so that each neighbourhood $L(\boldsymbol{x}, P_i)$ contains exactly k points from the i-th group of the training set. Then the classifier coming from (3.17) has a simple form:

$$d(\boldsymbol{x}) = \arg \min_{i=1,\ldots,K} \widehat{\lambda}_d(L(\boldsymbol{x}, P_i)), \tag{3.18}$$

where $L(\boldsymbol{x}, P_i)$ is the distributional neighbourhood of $\boldsymbol{x}$ which contains exactly k points from the i-th group of the training set. As far as the levelsets of density are convex, the volume of distributional neighborhood can be estimated quite easily.

It is clear that the classifier depends on the choice of constant k. This choice should follow general rules used in classical k-nearest-neighbour method. Thus k should increase to infinity as n increases to infinity. On the other hand the proportion of k/n might be smaller for large values of n, formally k/n approaches zero as n goes to infinity.

We show that the classifier (3.18) is meaningful and applicable when considering elliptically symmetric distributions.

Let us consider two elliptically symmetric distributions P_1 and P_2 on $\mathbb{R}^d$ with densities $f_1(\cdot)$ and $f_2(\cdot)$. Assume:

(Q1) $f_j(\boldsymbol{x}) = \frac{c}{|\boldsymbol{\Sigma}_j|^{1/2}} g\left(M_j(\boldsymbol{x})\right)$, where $M_j(\boldsymbol{x}) = \left[(\boldsymbol{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)\right]^{1/2}$ denote Mahalanobis distances of $\boldsymbol{x}$ from the P_j, $j = 1, 2$, and c is normalizing constant,

(Q2) g be continuous function, for which $g(cx) < g(x)$ for arbitrary $x \in \mathbb{R}^+$ and $c > 1$,

(Q3) distribution P be a mixture of P_1 and P_2: $P = \pi_1 P_1 + \pi_2 P_2$.

Consider two sequences of integers $\{n_i\}_{i\in\mathbb{N}}$ and $\{k_i\}_{i\in\mathbb{N}}$ such that: $\lim_{i\to\infty} n_i = \infty$, $\lim_{i\to\infty} k_i = \infty$ and $\lim_{i\to\infty} k_i/n_i = 0$.

We consider a sequence of independent d-dimensional random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ with the same distribution P. For any fixed $i \in \mathbb{N}$ and any fixed $\boldsymbol{x} \in \mathbb{R}^d$ we can divide the random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_i}$ into two parts:

- points that come from P_1 can be sorted according to their Mahalanobis distance to P_1 in an increasing order: $\boldsymbol{X}_{1:1} \prec \boldsymbol{X}_{2:1} \prec \ldots \prec \boldsymbol{X}_{m_1(i):1} \prec \boldsymbol{x} \prec \boldsymbol{X}_{m_1(i)+1:1} \prec \ldots \prec \boldsymbol{X}_{m_1(i)+k_i:1} \prec \ldots \prec \boldsymbol{X}_{u_i:1}$,
- points that come from P_2 can be sorted according to their Mahalanobis distance to P_2 in an increasing order: $\boldsymbol{X}_{1:2} \prec \boldsymbol{X}_{2:2} \prec \ldots \prec \boldsymbol{X}_{m_2(i):2} \prec \boldsymbol{x} \prec \boldsymbol{X}_{m_2(i)+1:2} \prec \ldots \prec \boldsymbol{X}_{m_2(i)+k_i:2} \prec \ldots \prec \boldsymbol{X}_{v_i:2}$,

Notice that the same ordering can be based on the halfspace depth (or any other depth function in the sense of the Definition 1.1). This important fact can be seen for example from Lemma 3.4.

In what follows we use neighbourhood of a fixed point $\boldsymbol{x} \in \mathbb{R}^d$ (whose Mahalanobis depth is denoted by $M_j := M_j(\boldsymbol{x})$, $j = 1, 2$, for simplicity) defined as

$$O_j(h) := \left\{\boldsymbol{y} \in \mathbb{R}^d : M_j(\boldsymbol{y}) \in [M_j, M_j + h]\right\}, j = 1, 2.$$

Recall that $\lambda_d(O_j(h)) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \left|\boldsymbol{\Sigma}_j\right|^{1/2} \left[(M_j + h)^d - M_j^d\right]$.

Theorem 3.5 *Consider the mixture of two distributions P (see Q1 - Q3 above), sequences of integers $\{n_i\}_{i\in\mathbb{N}}$ and $\{k_i\}_{i\in\mathbb{N}}$ and ordered random samples as described above. Let $\boldsymbol{x}$ be any fixed point in $\mathbb{R}^d$. For any $i \in \mathbb{N}$ define a random variables $C_1(i) := M_1(\boldsymbol{X}_{m_1(i)+k_i:1}) - M_1$ and $C_2(i) := M_2(\boldsymbol{X}_{m_2(i)+k_i:2}) - M_2$. In the considered situation it holds:*

$$\frac{\lambda_d\left(O_1\left(C_1\left(i\right)\right)\right)}{\lambda_d\left(O_2\left(C_2\left(i\right)\right)\right)} \to \frac{\pi_2 f_2(\boldsymbol{x})}{\pi_1 f_1(\boldsymbol{x})} \quad \textit{in probability.}$$

The theorem implies asymptotical equivalence of the classifier (3.18) and the Bayes classifier (3.1) under the assumption that the estimates $\widehat{\lambda}_d(L(\boldsymbol{x}, P_i))$ are consistent.

Proof: We want to show that

$$\forall \epsilon > 0\ \exists i_0 \in \mathbb{N} :\ i \geq i_0 \Rightarrow \mathrm{P}\left(\left|\frac{\lambda_d\left(O_1\left(C_1\left(i\right)\right)\right)}{\lambda_d\left(O_2\left(C_2\left(i\right)\right)\right)} \Big/ \frac{\pi_2 f_2(\boldsymbol{x})}{\pi_1 f_1(\boldsymbol{x})} - 1\right| > \epsilon\right) < \epsilon. \quad (3.19)$$

For any given $\epsilon > 0$ we can find constant $c_0(\epsilon) > 0$ such that $\frac{g(M_j + c_0(\epsilon))}{g(M_j)} < 1 + \epsilon$ for both $j = 1, 2$. Notice that this inequality implies $\frac{g(M_j+c)}{g(M_j)} < 1 + \epsilon$ for both $j = 1, 2$ for all $c \in [0, c_0(\epsilon)]$. Denote $p(\epsilon) := \min\left\{\pi_1 \mathrm{P}_1\left(O_1(c_0(\epsilon))\right), \pi_2 \mathrm{P}_2\left(O_2(c_0(\epsilon))\right)\right\}$.

Assume $i \in \mathbb{N}$ to be large enough to ensure

(A1) $k_i/n_i < p(\epsilon)/2$ and

(A2) $k_i^{-1/4} < \epsilon$.

In the three following steps we show that for any $i \in \mathbb{N}$ satisfying these two assumptions the inequality in (3.19) holds. Since now assume i to be fixed (satisfying conditions above) and we write k, n and $C_j, j = 1, 2$ instead of k_i, n_i and $C_j(i), j = 1, 2$ for simplicity.

Step 1:
We can find positive (uniquely determined) constants c_1 and c_2 such that

$$\pi_1 \mathrm{P}_1\left(O_1(c_1)\right) = k/n = \pi_2 \mathrm{P}_2\left(O_2(c_2)\right). \tag{3.20}$$

Obviously $0 < c_j < c_0$ for both $j = 1, 2$.

Now

$$\frac{\frac{\lambda_d(O_1(c_1))}{\lambda_d(O_2(c_2))}}{\frac{\pi_2 f_2(\boldsymbol{x})}{\pi_1 f_1(\boldsymbol{x})}} = \frac{\frac{|\boldsymbol{\Sigma}_1|^{1/2}\left[(M_1+c_1)^d - M_1^d\right]}{|\boldsymbol{\Sigma}_2|^{1/2}\left[(M_2+c_2)^d - M_2^d\right]}}{\frac{\pi_2 |\boldsymbol{\Sigma}_2|^{-1/2} g(M_2)}{\pi_1 |\boldsymbol{\Sigma}_1|^{-1/2} g(M_1)}} = \frac{\pi_1 g(M_1)\left[(M_1+c_1)^d - M_1^d\right]}{\pi_2 g(M_2)\left[(M_2+c_2)^d - M_2^d\right]}. \tag{3.21}$$

Using the equation (3.20) the ratio (3.21) can be written as

$$\begin{aligned} &\frac{\pi_2 \mathrm{P}_2(O_2(c_2))}{\pi_2 g(M_2)\left[(M_2+c_2)^d - M_2^d\right]} \cdot \frac{\pi_1 g(M_1)\left[(M_1+c_1)^d - M_1^d\right]}{\pi_1 \mathrm{P}_1(O_1(c_1))} = \\ &= \frac{\int_{M_2}^{M_2+c_2} g(r) r^{d-1} \mathrm{d}r}{g(M_2)\left[(M_2+c_2)^d - M_2^d\right]} \cdot \frac{g(M_1)\left[(M_1+c_1)^d - M_1^d\right]}{\int_{M_1}^{M_1+c_1} g(r) r^{d-1} \mathrm{d}r}. \end{aligned} \tag{3.22}$$

Now we can find upper bound for this ratio (and analogous lower bound). Since $g(\cdot)$ is decreasing function, it holds $g(r) < g(M_2 + c_2)$ for all $r \in [M_2, M_2 + c_2)$ and $g(r) > g(M_1)$ for all $r \in (M_1, M_1 + c_1)$. Hence (3.22) is bounded above by

$$\frac{\int_{M_2}^{M_2+c_2} g(M_2+c_2) r^{d-1} \mathrm{d}r}{g(M_2)\left[(M_2+c_2)^d - M_2^d\right]} \cdot \frac{g(M_1)\left[(M_1+c_1)^d - M_1^d\right]}{\int_{M_1}^{M_1+c_1} g(M_1) r^{d-1} \mathrm{d}r} = \frac{g(M_2+c_2)}{g(M_2)} \cdot \frac{g(M_1)}{g(M_1)} < 1+\epsilon.$$

Similarly, the lower bound for the ratio can be computed.

Step 2:
We find positive constants c_1^L, c_1^U and c_2^L, c_2^U such that

$$\begin{aligned} \pi_1 \mathrm{P}_1\left(O_1(c_1^L)\right) &= \tfrac{k-k^{3/4}}{n} = \pi_2 \mathrm{P}_2\left(O_2(c_2^L)\right), \\ \pi_1 \mathrm{P}_1\left(O_1(c_1^U)\right) &= \tfrac{k+k^{3/4}}{n} = \pi_2 \mathrm{P}_2\left(O_2(c_2^U)\right). \end{aligned}$$

These constants are again unique and less than c_0.

Now consider any constant $c_1^S \in [c_1^L, c_1^U]$ and $c_2^S \in [c_2^L, c_2^U]$. We do not assume $\pi_1 \mathrm{P}_1\left(O_1(c_1^S)\right) = \pi_2 \mathrm{P}_2\left(O_2(c_2^S)\right)$. Nevertheless, it can proved that the ratio $\frac{\lambda_d\left(O_1(c_1^S)\right)}{\lambda_d\left(O_2(c_2^S)\right)} / \frac{\pi_2 f_2(\boldsymbol{x})}{\pi_1 f_1(\boldsymbol{x})}$ is close to one.

We can proceed similarly as in the first step:

$$\frac{\lambda_d\left(O_1\left(c_1^S\right)\right)}{\lambda_d\left(O_2\left(c_2^S\right)\right)}\Big/\frac{\pi_2 f_2(\boldsymbol{x})}{\pi_1 f_1(\boldsymbol{x})} = \frac{\pi_1 g(M_1)\left[(M_1+c_1^S)^d - M_1^d\right]}{\pi_2 g(M_2)\left[(M_2+c_2^S)^d - M_2^d\right]}. \tag{3.23}$$

The fraction can be extended by

$$\frac{\pi_1 P_1(O_1(c_1^S))}{\pi_1 P_1(O_1(c_1^S))}\frac{\pi_2 P_2(O_2(c_2^S))}{\pi_2 P_2(O_2(c_2^S))}\frac{\pi_2 P_2(O_2(c_2))}{\pi_1 P_1(O_1(c_1))},$$

where the last term is equal to one from (3.20). After a convenient arrangement we get (3.23) equals to

$$\frac{g(M_1)\left[(M_1+c_1^S)^d - M_1^d\right]}{P_1(O_1(c_1^S))}\cdot\frac{P_1(O_1(c_1^S))}{P_1(O_1(c_1))}\cdot\frac{P_2(O_2(c_2^S))}{g(M_2)\left[(M_2+c_2^S)^d - M_2^d\right]}\cdot\frac{P_2(O_2(c_2))}{P_2(O_2(c_2^S))}.$$

Ratios $\frac{\mathrm{P}(O_j(c_j^S))}{\mathrm{P}(O_j(c_j))}$ are not greater than $\frac{(k+k^{3/4})/n}{k/n} = 1+k^{-1/4}$ and not smaller than $\frac{(k-k^{3/4})/n}{k/n} = 1-k^{-1/4}$. Recall that k is so big that $k^{-1/4} < \epsilon$. The first and the third term are both bounded similarly as the ratio in the step 1.

The considered ratio is thus not greater than $\frac{(1+\epsilon)^2}{1-\epsilon}$ and not less than $\frac{1-\epsilon}{(1+\epsilon)^2}$.

Step 3:
We show that $C_j \in [c_j^L, c_j^U]$ with probability greater than $1-2\epsilon$ both for $j=1,2$. Consider a random sample of n points from the mixture P (some of the randomly sampled points are from P_1 and some are from P_2).

Let $Z_j^L, j=1,2$, denote numbers of points from P_j lying in $O_j(c_j^L)$. $Z_j^L, j=1,2$, are binomial random variables: $Z_j^L \sim \mathrm{Bi}\left(\frac{k-k^{3/4}}{n}, n\right)$. Let $Z_j^U, j=1,2$, denote numbers of points from P_j lying in $O_j(c_j^U)$. $Z_j^U, j=1,2$, are binomial random variables: $Z_j^U \sim \mathrm{Bi}\left(\frac{k+k^{3/4}}{n}, n\right)$.

Obviously $C_j \notin [c_j^L, c_j^U]$ iff either $Z_j^L > k$ (in that case $C_j < c_j^L$) or $Z_j^U < k$ (in that case $C_j > c_j^U$). Now

$$\mathrm{P}\left(Z_j^L > k\right) = \mathrm{P}\left(\frac{Z_j^L - EZ_j^L}{SD(Z_j^L)} > \frac{k-(k-k^{3/4})}{\sqrt{(k-k^{3/4})(1-\frac{k-k^{3/4}}{n})}}\right).$$

The standard deviation of the considered binomial distribution is smaller than $k^{1/2}$, hence

$$\mathrm{P}\left(Z_j^L > k\right) < \mathrm{P}\left(\frac{Z_j^L - EZ_j^L}{SD(Z_j^L)} > k^{1/4}\right) \le k^{-1/2} < \epsilon,$$

where the second inequality follows from the Chebyshev's inequality and the last inequality follows from the assumption (A2).

Similarly it can be shown that $\mathrm{P}\left(Z_j^U < k\right) < \epsilon$. Hence $P\left(C_j \in [c_j^L, c_j^U]\right) > 1-2\epsilon$.

□

Newly proposed modified k-nearest-neighbour method is an alternative to the methods based on density estimation presented in the Section 3.4.2. In contrast to these methods, modified k-NN does not need kernel density estimation. Its implementation is quite easy. While classical k-NN method suffers from curse of dimensionality, modified k-NN prevent the problem by using alternatively understood notion of neighbourhood. In contrast to the maximal depth classifier and related classifiers, modified k-NN method does not have problems with classification when the considered distributions differ in dispersion.

3.5 Alternative depth-based method of discrimination

An alternative approach to the two-class discrimination problem based on data depth was recently proposed by Li, Cuesta-Albertos and Liu [34]. Their classifier is based on so called DD–plot, a visual tool where depth of point with respect to one distribution (say P_1) is plotted against its depth with respect to other distribution (P_2). Formally:

Definition 3.5 *Denote* $\boldsymbol{X}^{(1)} := \boldsymbol{X}_{1,1}, \ldots, \boldsymbol{X}_{1,n_1}$ *a random sample from* P_1 *and* $\boldsymbol{X}^{(2)} := \boldsymbol{X}_{2,1}, \ldots, \boldsymbol{X}_{2,n_2}$ *the random sample from* P_2*. The DD-plot is defined by*

$$DD(P_1, P_2) = \left\{ [D(\boldsymbol{x}; P_1), D(\boldsymbol{x}; P_2)], \boldsymbol{x} \in \boldsymbol{X}^{(1)} \cup \boldsymbol{X}^{(2)} \right\}.$$

When P_1 *and* P_2 *are unknown, the DD-plot is defined as*

$$DD(\hat{P}_1, \hat{P}_2) = \left\{ [D(\boldsymbol{x}; \hat{P}_1), D(\boldsymbol{x}; \hat{P}_2)], \boldsymbol{x} \in \boldsymbol{X}^{(1)} \cup \boldsymbol{X}^{(2)} \right\}.$$

The classifier is constructed by finding a curve best separating the two samples in the DD–plot. Li, Cuesta-Albertos and Liu consider a polynomial function separating the two samples. For example, consider the straight line which separates the two samples. This line should go through the point [0,0]. The classifier thus has the following form:

$$\begin{aligned} D(\boldsymbol{x}; \hat{P}_2) > \widehat{k} D(\boldsymbol{x}; \hat{P}_1) &\implies d(\boldsymbol{x}) = 2 \\ D(\boldsymbol{x}; \hat{P}_2) < \widehat{k} D(\boldsymbol{x}; \hat{P}_1) &\implies d(\boldsymbol{x}) = 1, \end{aligned} \tag{3.24}$$

where $\widehat{k}$ is estimated slope of the separating line. $\widehat{k}$ is chosen such that empirical misclassification rate is made minimal: $\widehat{k} = \arg\min_k \widehat{\Delta}(k)$, where

$$\widehat{\Delta}(k) = \widehat{\pi}_1 \frac{1}{n_1} \sum_{i=1}^{n_1} I_{[D(\boldsymbol{X}_{1,i}; \hat{P}_2) > kD(\boldsymbol{X}_{1,i}; \hat{P}_1)]} + \widehat{\pi}_2 \frac{1}{n_2} \sum_{j=1}^{n_2} I_{[D(\boldsymbol{X}_{2,j}; \hat{P}_2) < kD(\boldsymbol{X}_{2,j}; \hat{P}_2)]}.$$

The minimization is made over the set of $n_1 + n_2$ possible slopes determined by points of the training set in practice. The classifier is proved to be asymptotically equivalent to the Bayes rule under the same conditions as the maximal depth classifier (3.3). Moreover, it performs quite well (though not optimal) in the case of unequal dispersions:

Consider the case of two bivariate normal distributions described in Section 3.3.1. We consider a training set of 1000 observations from each of the two distributions.

Halfspace depth of the points from the training set is estimated and the DD-plot and classifier based on the DD-plot are constructed. The DD-plot in Figure 3.5 contains only 100 points from each group for simplicity. It can be easily seen that there is only a small number of points, whose depth with respect to $\hat{P}_2$ is greater than their depth with respect to $\hat{P}_1$. Hence the maximal depth classifier classifies majority of new observations to the group 1 (with higher dispersion). The average misclassification rate is thus high. In contrary, classifier (3.24) uses other dividing line - the line whose slope is about 0.135. The average misclassification rate is then close to the optimal Bayes risk.

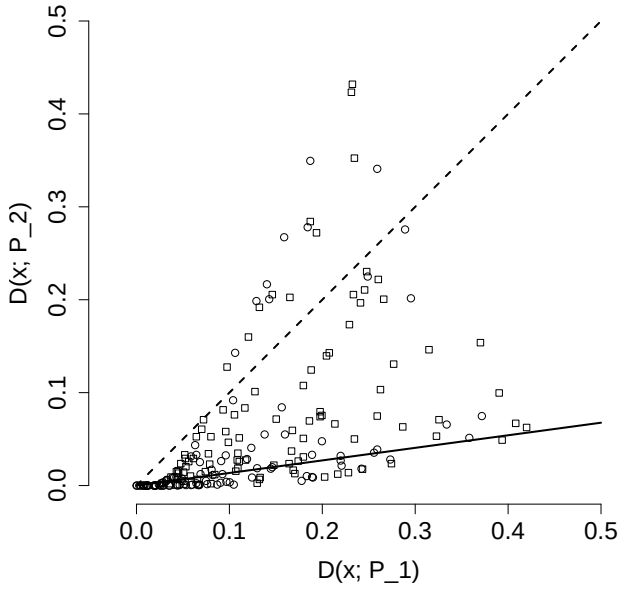


Figure 3.5: DD-plot of two normal distributions: points from P_1 are marked by circles, points from P_2 by squares.

The classifier based on DD-plot uses more general idea of dimension reduction. The idea consists in the reduction of the data dimensionality and solving the problem in the lower dimensions. Scott in [48] wrote: "Multivariate data in $\mathbb{R}^d$ are almost never d-dimensional. That is, the underlying structure of data in $\mathbb{R}^d$ is almost always of dimension lower than d." Hence a suitable transformation of the data $f : \mathbb{R}^d \to \mathbb{R}^{d'}$, where $d' < d$ is needed. A possible method for dimension reduction in computing the data depth. Any depth function transforms $\mathbb{R}^d \to \mathbb{R}^+$. The applicability of this approach is limited by high computation costs needed for depth computation (as discussed in Chapter 1). Thus we can use this approach if $d = 2, d = 3, d = 4$, but not more than $d = 5$ for majority of known affine invariant depth functions. When the dimension of original data is higher, the methodology of data depth could be combined for example with principal component analysis - a well known method for reduction of dimensionality.

3.6 Observations of zero depth

Most of the depth functions assign zero depth to points outside the convex hull of the support of distribution. It means that points outside the convex hull of a training set have empirical depth equal to zero.

Let us consider a situation that for a new observation $\boldsymbol{x}$ it holds $D(\boldsymbol{x}; \hat{P}_i) = 0$, for all $i = 1, \ldots, K$ - the observation has zero depth with respect to all groups of points from training set. There are several possibilities how to solve this problem:

1. Some depth function, which is positive for any point, may be used. Mahalanobis depth or L_1–depth are convenient in this case. This method was used for example by Mosler and Hoberg [29] who combined zonoid and Mahalanobis depth.

2. New observation can be classified to the group, from whose deepest point it has minimal Euclidean distance.

3. New observation can be classified to the group, from whose specific central region it has minimal Euclidean distance.

3.7 Simulation study

We illustrate advantages and weak points of several classifiers in simulation study. We compare two traditional classifiers to some classifiers based on data depth. Linear discriminant analysis (LDA) represents traditional parametric approach. Nonparametric approach is represented by k-nearest-neighbour method (k-NN). Depth based classifiers are represented by the maximal depth classifier based on adjusted outlyingness introduced in Section 3.3.2, classifier based on DD-plot (where halfspace depth is considered) introduced in Section 3.5, and newly proposed modified k-nearest-neighbour method introduced in Section 3.4.3. In all simulations we used $k = 51$ (we considered 51 nearest neighbours).

We deal with four bivariate situations considered in this chapter: two normal distributions differ in location only (the location-shift model), two normal distributions differ in location and dispersion, and normal and skewed normal distribution (example used in [25]). The fourth simulation illustrates an advantage gain by use of the weighted data depth instead of the halfspace depth. Two uniform distributions, whose disjoint nonconvex supports together form a unit circle, are considered here.

A training set of 2000 points was generated. Subsequently, classifiers were constructed. Finally, the classifiers were "tested" on a test set of another 2000 points, that is the average misclassification rates were estimated on the independent sample of points.

We considered three situations corresponding to the three different prior probabilities. First, equal prior probabilities were considered. In this situation 1000 points of a training set was generated from each group. The test set was also divided between the two groups in 1:1 ratio. Later, we considered 3:1 ratio of points, generating 1500 points from group 1 and 500 points from group 2 (both for training set and test set). Finally, reverse ratio of points was considered (that is 500 points from group 1 and 1500 points from group 2).

For each situation the simulation was repeated 100 times.

3.7.1 Two normal distributions differ in location only

Let us consider two bivariate normal distributions which differ only in location:

$$P_1 = N\left(\mathbf{0}, \mathbf{I}\right), P_2 = N\left((2,0)^T, \mathbf{I}\right).$$

The Bayes rule in this case has the following form:

$$d(\boldsymbol{x} = (x_1, x_2)^T) = 1 \text{ iff } x_1 < 1 - \frac{1}{2}\ln\left(\frac{\pi_2}{\pi_1}\right).$$

Corresponding minimal error rates (optimal Bayes risks) are evaluated in the first row of the Table 3.1. The remaining rows include estimated average misclassification rates (AMR) for particular classifiers.

	$\pi_1 = \pi_2 = 0.5$	$\pi_1 = 0.75; \pi_2 = 0.25$	$\pi_1 = 0.25; \pi_2 = 0.75$
Bayes	*0.1587*	*0.1270*	*0.1270*
LDA	0.1588 (0.0069)	0.1270 (0.0065)	0.1265 (0.0064)
k-NN	0.1612 (0.0078)	0.1293 (0.0067)	0.1287 (0.0070)
max. depth	0.1611 (0.0073)	0.1532 (0.0117)	0.1537 (0.0116)
DD-plot	0.1600 (0.0078)	0.1289 (0.0075)	0.1284 (0.0073)
modified k-NN	0.1582 (0.0073)	0.1266 (0.0065)	0.1263 (0.0069)

Table 3.1: Estimated average misclassification rates of different classifiers (estimated standard deviations of AMR estimates are given parenthesis).

It can be seen from Table 3.1 that all considered classifiers works quite well when equal prior probabilities are considered. Their AMRs are near the optimal Bayes risk. When unequal priors are considered, the maximal depth classifier has significantly higher AMR. The other classifiers are again near the optimal classifier.

More detailed information about error rates is displayed in Table 3.2. Error rates for group 1 (written in left part of each cell) are compared to the error rates for group 2 (written in the right part of each cell).

	$\pi_1 = \pi_2 = 0.5$	$\pi_1 = 0.75; \pi_2 = 0.25$	$\pi_1 = 0.25; \pi_2 = 0.75$
Bayes	*0.1587 ; 0.1587*	*0.0607 ; 0.3261*	*0.3261 ; 0.0607*
LDA	0.1580 ; 0.1596	0.0609 ; 0.3255	0.3225 ; 0.0612
k-NN	0.1583 ; 0.1640	0.0594 ; 0.3390	0.3376 ; 0.0590
max. depth	0.1605 ; 0.1617	0.1438 ; 0.1811	0.1815 ; 0.1444
DD-plot	0.1571 ; 0.1629	0.0618 ; 0.3304	0.3258 ; 0.0625
modified k-NN	0.1574 ; 0.1590	0.0622 ; 0.3196	0.3197 ; 0.0619

Table 3.2: Misclassification rates for group 1 (left) and group 2 (right).

Different behaviour of the maximal depth classifier can be seen from Table 3.2. Its error rates do not differ as much as the error rates of the other classifiers when considered unequal prior probabilities.

3.7.2 Two normal distributions differ in location and dispersion

Let us consider the case of two bivariate normal distributions which differ location and dispersion, introduced in Section 3.3.1:

$$P_1 = N\left(\mathbf{0}, 4\mathbf{I}\right), P_2 = N\left((1,0)^T, \mathbf{I}\right).$$

The Bayes rule in this case has the following form:

$$d(\boldsymbol{x} = (x_1, x_2)^T) = 1 \text{ iff } \left(x_1 - \frac{4}{3}\right)^2 + (x_2)^2 - \frac{8}{3}\ln\left(\frac{\pi_2}{\pi_1}\right) - \frac{4}{9} > 0.$$

Similar tables as in the previous case are provided: Table 3.3 summarizes (estimated) average misclassification rates for particular classifiers, and Table 3.4 provides information about error rates in individual groups.

	$\pi_1 = \pi_2 = 0.5$	$\pi_1 = 0.75; \pi_2 = 0.25$	$\pi_1 = 0.25; \pi_2 = 0.75$
Bayes	*0.2408*	*0.2267*	*0.1539*
LDA	0.3561 (0.0110)	0.2664 (0.0057)	0.1959 (0.0058)
k-NN	0.2472 (0.0094)	0.2342 (0.0083)	0.1661 (0.0066)
max. depth	0.4242 (0.0169)	0.2357 (0.0062)	0.5812 (0.0484)
DD-plot	0.2477 (0.0087)	0.2286 (0.0085)	0.1589 (0.0073)
modified k-NN	0.2448 (0.0094)	0.2293 (0.0078)	0.1567 (0.0075)

Table 3.3: Estimated average misclassification rates of different classifiers (estimated standard deviations of AMR estimates are given parenthesis).

	$\pi_1 = \pi_2 = 0.5$	$\pi_1 = 0.75; \pi_2 = 0.25$	$\pi_1 = 0.25; \pi_2 = 0.75$
Bayes	*0.3409 ; 0.1406*	*0.1144 ; 0.5637*	*0.5105 ; 0.0350*
LDA	0.4027 ; 0.3096	0.0225 ; 0.9982	0.7574 ; 0.0088
k-NN	0.3611 ; 0.1333	0.1137 ; 0.5958	0.6057 ; 0.0196
max. depth	0.0463 ; 0.8021	0.0404 ; 0.8216	0.0627 ; 0.7540
DD-plot	0.3454 ; 0.1499	0.1100 ; 0.5843	0.5234 ; 0.0374
modified k-NN	0.3478 ; 0.1419	0.1178 ; 0.5638	0.5091 ; 0.0392

Table 3.4: Misclassification rates for group 1 (left) and group 2 (right).

It can be seen from Table 3.3, that LDA and maximal depth classifiers are not appropriate in this case even if the priors are equal. All three remaining classifiers perform quite well. Classifiers based on data depth (DD-plot classifier and modified k-NN classifier) seem to outperform classical k-NN slightly when priors are not equal.

Table 3.4 documents poor behaviour of the maximal depth classifier in this case. This classifier clearly "prefers" group 1 and hence it has high error rate for group 2 regardless of the prior probabilities. This leads to high average misclassification rate if $\pi_2 >> \pi_1$. For example AMR is equal to 0.5812 if $\pi_2 = 0.75$, as can be seen in Table 3.3.

3.7.3 Normal and skewed normal distribution

Let us recall the example from Section 3.3.2. A normal and a skewed normal distribution is considered:

$$P_1 = N(\mathbf{0}, \mathbf{I}), P_2 = SN((-2,-2)^T, \mathbf{I}, (5,5)^T).$$

The Bayes rule can be obtained by solving the equation $\pi_1 f_1(\boldsymbol{x}) = \pi_2 f_2(\boldsymbol{x})$ in $\boldsymbol{x} = (x_1, x_2)^T$. Using a symbol $\phi(\cdot)$ for the density function of the standard normal distribution and a symbol $\Phi(\cdot)$ for its cumulative distribution function, we have

$$\begin{aligned} \pi_1 f_1(x) &= \pi_2 f_2(x), \\ \pi_1 \cdot \phi(x_1)\phi(x_2) &= \pi_2 \cdot 2\phi(x_1+2)\phi(x_2+2)\Phi(5(x_1+2+x_2+2)), \\ \frac{\pi_1 \phi(x_1)\phi(x_2)}{2\pi_2 \phi(x_1+2)\phi(x_2+2)} &= \Phi(5(x_1+2+x_2+2)). \end{aligned}$$

We can seek solutions on straight lines of the form $x_2 = -x_1 + q$, where q is some unknown constant.

$$\frac{\pi_1 \phi(x_1)\phi(-x_1+q)}{2\pi_2 \phi(x_1+2)\phi(-x_1+q+2)} = \Phi(5(x_1+2-x_1+q+2)).$$

Now we utilize the known form of the density function ϕ: $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. We get the following equation:

$$\frac{\pi_1}{2\pi_2} e^{2(q+2)} = \Phi\left(5(q+4)\right). \tag{3.25}$$

This formula does not include x_1 term, hence the dividing surfaces has the form of lines $x_2 = -x_1 + q$ for some values of unknown constant q. Equation (3.25) can be solved numerically. There is one possible solution in the interval (-5,-4) and one in the interval (-3,-1) for all three considered prior distributions.

It should be verified that there are no more solutions. This could be done by differentiating (3.25): equation $2e^{2(q+2)}\pi_1/\pi_2 = 2\frac{1}{\sqrt{2\pi}} e^{-(5q+20)^2/2}$ leads to a quadratic equation with two real roots. Hence it is proved that there can not be more than two real solutions of (3.25).

The Bayes rule has the following form:

$$d(\boldsymbol{x} = (x_1, x_2)^T) = 1 \text{ iff } (x_2 > -x_1 + q_2 \text{ or } x_2 < -x_1 + q_1),$$

where

$$\begin{aligned} q_1 = -4.55 \text{ and } q_2 = -1.65 &\quad \text{if } \pi_1 = 0.50, \\ q_1 = -4.46 \text{ and } q_2 = -2.20 &\quad \text{if } \pi_1 = 0.75, \\ q_1 = -4.63 \text{ and } q_2 = -1.10 &\quad \text{if } \pi_1 = 0.25. \end{aligned}$$

Remind that the values of q_1 and q_2 are only approximate. For $\pi_1 = \pi_2$, a misclassification rate for the group 1 can be expressed as $\Phi(q_2/\sqrt{2}) - \Phi(q_1/\sqrt{2})$ and a misclassification rate for the group 2 can be find out by simulation.

Results of the simulations are summarized in Tables 3.3 and 3.4. It can be seen that all considered classifiers perform quite well when equal priors are assumed. Maximal depth classifier has higher average misclassification rates when considered unequal priors. This is rather disappointing fact, because the classifier is based on adjusted outlyingness and hence it is constructed right for the case of skewed distributions. Unfortunately, unequal priors corrupt its behaviour.

	$\pi_1 = \pi_2 = 0.5$	$\pi_1 = 0.75; \pi_2 = 0.25$	$\pi_1 = 0.25; \pi_2 = 0.75$
Bayes	*0.1089*	*0.0952*	*0.0847*
LDA	0.1098 (0.0065)	0.0950 (0.0068)	0.0851 (0.0063)
k-NN	0.1100 (0.0062)	0.0962 (0.0068)	0.0865 (0.0062)
max. depth	0.1108 (0.0065)	0.1112 (0.0089)	0.1032 (0.0106)
DD-plot	0.1096 (0.0068)	0.0958 (0.0068)	0.0859 (0.0067)
modified k-NN	0.1087 (0.0067)	0.0941 (0.0066)	0.0849 (0.0063)

Table 3.5: Estimated average misclassification rates of different classifiers (estimated standard deviations of AMR estimates are given parenthesis).

	$\pi_1 = \pi_2 = 0.5$	$\pi_1 = 0.75; \pi_2 = 0.25$	$\pi_1 = 0.25; \pi_2 = 0.75$
Bayes	*0.1205 ; 0.0972*	*0.0589 ; 0.2043*	*0.2169 ; 0.0407*
LDA	0.1394 ; 0.0803	0.0616 ; 0.1949	0.2292 ; 0.0370
k-NN	0.1237 ; 0.0962	0.0566 ; 0.2149	0.2316 ; 0.0381
max. depth	0.1180 ; 0.1036	0.1066 ; 0.1248	0.1312 ; 0.0939
DD-plot	0.1191 ; 0.1002	0.0589 ; 0.2064	0.2149 ; 0.0429
modified k-NN	0.1311 ; 0.0862	0.0613 ; 0.1922	0.2343 ; 0.0351

Table 3.6: Misclassification rates for group 1 (left) and group 2 (right).

3.7.4 Two uniform distributions on disjoint nonconvex supports

We have already seen that DD-plot classifier and modified k-NN classifier work quite well in the situation of two unimodal distributions with convex levelsets of density. Now we consider the case of two uniform distributions with nonconvex supports. The supports are disjoint, hence the Bayes risk is equal to zero (perfect separation exists). Supports of the considered distributions, whose union forms the unit circle, can be seen in Figure 3.6. Formally, we consider P_1 and P_2 uniform distributions with the following supports:

- $\mathrm{sp}(P_1) = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 < 1\} \cap \{(x_1, x_2) \in \mathbb{R}^2 : x_1/x_2 \in (-1, 0) \cup (1, \infty)\}$,
- $\mathrm{sp}(P_2) = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 < 1\} \setminus \mathrm{sp}(P_1)$.

The example should illustrate how much it is important to choose a convenient depth function when using a depth based classifier. We used the halfspace depth and the band depth proposed in Section 2.2 (with parameter $h = 0.2$).

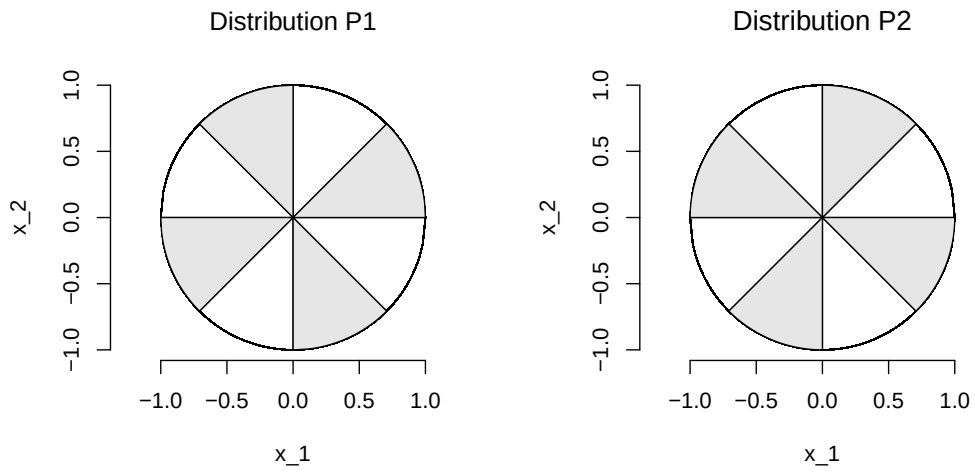


Figure 3.6: Supports of P_1 (left) and P_2 (right) dark coloured.

Recall the example of a uniform distribution on square-shaped support introduced in Section 2.7.1. We observed that the central regions of the band depth correspond to the shape of the support "better" than the central regions of the halfspace depth. Figure 3.7 compares central regions of the band depth and the halfspace depth of P_1.

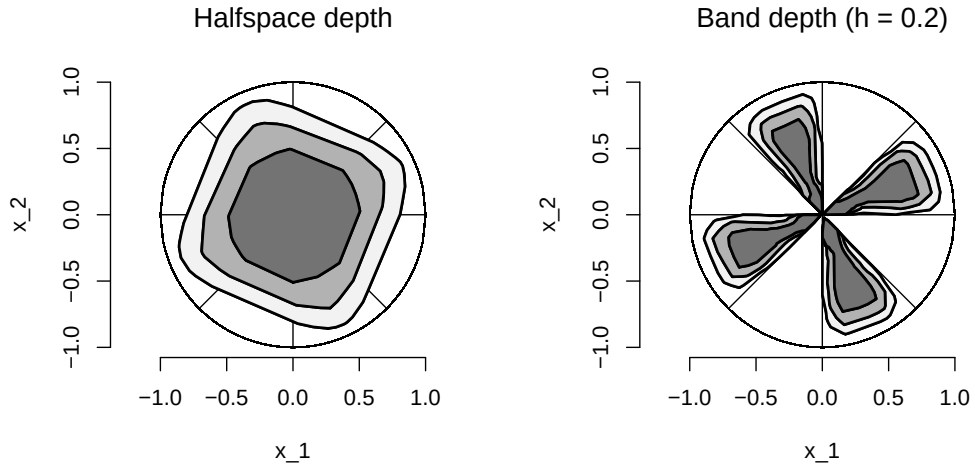


Figure 3.7: Areas of 25%, 50% and 75% of deepest points with respect to P_1 considering the halfspace depth (left) and the band depth with $h = 0.2$ (right).

The correspondence between shapes of central regions and the shape of the distribution is very important for discrimination. This fact is apparent when comparing misclassification rates shown in Table 3.7. For example, in the case $\pi_1 = \pi_2 = 1/2$ the DD-plot classifier has average misclassification rate about 0.26 when using the halfspace depth, but only about 0.07 when using the band depth (with $h = 0.2$). In the same case, the modified k-NN classifier has the average misclassification rate about 0.39 when using the halfspace depth, but only about 0.09 when using the band depth (with $h = 0.2$).

We can conclude that both advanced depth based classifiers perform much better when using band depth than the halfspace depth in this example.

	$\pi_1 = \pi_2 = 0.5$	$\pi_1 = 0.75; \pi_2 = 0.25$	$\pi_1 = 0.25; \pi_2 = 0.75$
LDA	0.5001 (0.0132)	0.2500 (0)	0.2500 (0)
k-NN	0.0490 (0.0067)	0.0947 (0.0063)	0.0951 (0.0062)
max. depth	0.7086 (0.0749)	0.6022 (0.1199)	0.6090 (0.1092)
DD-plot (*)	0.2593 (0.0139)	0.1478 (0.0109)	0.1470 (0.0094)
DD-plot (**)	0.0730 (0.0075)	0.0496 (0.0058)	0.0483 (0.0049)
modified k-NN (*)	0.3895 (0.0185)	0.2314 (0.0046)	0.2307 (0.0047)
modified k-NN (**)	0.0899 (0.0087)	0.0920 (0.0068)	0.0915 (0.0055)

Table 3.7: Estimated average misclassification rates of different classifiers (estimated standard deviations of AMR estimates are given parenthesis). Depth based classifiers use (*) the halfspace depth and (**) the band depth with $h = 0.2$.

Conclusion

In this work we introduced a new idea of the weighted data depth. This generalization of the well-known halfspace depth has several interesting properties. Mainly the possibility of nonconvex central regions is desirable in some applications. However, the weighted data depth is not affine invariant in general; only translation and rotation invariance can be proved.

Concept of weighted data depth provides a broad class of possible weight functions determined by chosen weight function. We specified several reasonable mild restrictions on the weight function to ensure a strong pointwise consistency of the weighted depth function. We also discussed restrictions which guarantee that the depth of points lying out of the probability support is equal to zero.

The advantage of possibly nonconvex central regions can be utilized for example in the discriminant analysis, as was shown at the end of the work.

In the part devoted to the use of data depth for discrimination purposes, we presented comprehensive critical review of depth-based methods proposed in the last ten years. We uncovered problems of standard depth-based classifiers when the considered distributions differ in dispersion. We introduced a new idea of modified k-nearest-neighbour classifier. Newly proposed classifier performs well in various situations. The method can be implemented more easily than comparable classifiers based on kernel density estimation. Together with DD-plot based classifier, it is probably the most promising depth-based classifier.

Besides the new ideas, this work provides a detailed review of existing depth functions. Weak and strong points of the data depth concept are discussed as well.

Bibliography

[1] Azzalini, A., Della Valle, A. The multivariate skew-normal distribution. *Biometrika*, 1996, vol. 83, no. 4, s. 715 – 726.

[2] Bárdossy, A., Singh, S. K. Robust estimation of hydrological model parameters. *Hydrology and Earth System Sciences*, 2008, vol. 12, s. 1273 – 1283.

[3] Billor, N., et al. Classification based on depth transvariations. *Journal of Classification*, 2008, vol. 25, s. 249 – 260.

[4] Bremner, D., Fukuda, K., Rosta, V. Primal-dual algorithms for data depth. In *Data depth: robust multivariate analysis, computational geometry and applications*. R. Y. Liu, R. Serfling, and D. L. Souvaine, eds. 1st edition. New York: American Mathematical Society, 2006. DIMACS series in discrete mathematics and theoretical computer science; 72. s. 171 – 194.

[5] Bremner, D., et al. Output-sensitive algorithms for Tukey depth and related problems. *Statistics and Computing*, 2008, vol. 18, s. 259 — 266.

[6] Brys, G., Hubert, M., Struyf, A. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 2004, vol. 13, no. 4, s. 996 – 1017.

[7] Cheng, A. Y., Quyang, M. On algorithms for simplicial depth. *Proceedings of the 13th Canadian Conference on Computational Geometry*, 2001, s. 53 — 56.

[8] Christmann, A., Rousseeuw, P. Measuring overlap in binary regression. *Computational Statistics and Data Analysis*, 2001, vol. 37, s. 65 – 75.

[9] Cramér, H., Wold, H. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1936, vol. 11, no. 4, s. 290 – 294.

[10] Donoho, D. L., Gasko, M. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, 1992, vol. 20, no. 4, s. 1803 – 1827.

[11] Dutta, S, Ghosh, A. K. On robust classification using projection depth. *Technical report no. R11/2009* [online]. Indian Statistical Institute. [cited 2011-07-26], available from <http://www.isical.ac.in/šatmath/html/publication/PD.pdf>.

[12] Dyckerhoff, R., Koshevoy, G. A., Mosler, K. Zonoid data depth: theory and computation. In *COMPSTAT 1996 - Proceedings in Computational Statistics: 12th Symposium held in Barcelona, Spain*. A. Prat editor. Heidelberg: Physica-Verlag, 1996. s. 235 – 240.

[13] Fang, K. T., Kotz, S., Ng, K. W. *Symmetric multivariate and related distributions*. 1st edition. London: Chapman and Hall, 1990. 220 s. Monographs on statistics and applied probability; 36. ISBN 0-412-31430-4.

[14] Fix, E., Hodges, J. L. Discriminatory analysis: nonparametric discrimination: consistency properties. *Technical report no. 4* [online]. Randolph Field, Texas: USAF School of Aviation Medicine, 1951. [cited 2011-07-27], available form <http://www.dtic.mil>.

[15] Fraiman, R., Liu, R. Y., Meloche, J. Multivariate density estimation by probing depth. In *L_1-statistical procedures and related topics*. Y. Dodge, editor. Hayward, CA: Institute of Mathematical Statistics, 1997. IMS Lecture notes– Monograph Series; 31. s. 415 – 430.

[16] Fukuda, K., Rosta, V. Data depth and optimization. *Technical Report* [online]. [cited 2010-07-01], available at <http://www.ifor.math.ethz.ch/about_us/press/Leitartikel_Marz_2005.pdf>.

[17] Garey, M. R., Johnson, D. S. *Computers and intractability: a guide to the theory of NP-completeness*. 1st edition. New York: W. H. Freeman and Company, 1979. 338 s. A series of books in the mathematical sciences. ISBN 0-7167-1045-5.

[18] Ghosh, A. K., Chaudhuri, P. On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 2005, vol. 32, s. 327 – 350.

[19] Hand, D. J. *Discrimination and Classification*. 1st edition. Chichester: Wiley, 1981. 218 s. Wiley series in probability and mathematical statistics. ISBN 0-471-28048-8.

[20] Hartikainen, A., Oja, H. On some parametric, nonparametric and semiparametric discrimination rules. In *Data depth: robust multivariate analysis, computational geometry and applications*. R. Y. Liu, R. Serfling, and D. L. Souvaine, eds. 1st edition. New York: American Mathematical Society, 2006. DIMACS series in discrete mathematics and theoretical computer science; 72. s. 61 – 70.

[21] Hassairi, A., Regaieg, O. On the Tukey depth of a continuous probability distribution. *Statistics and Probability Letters*, 2008, vol. 78, s. 2308 – 2313.

[22] Hlubinka, D., Kotík, L., Vencálek, O. Weighted data depth. *Kybernetika*, 2010, vol. 46, no. 1, s. 125 – 148.

[23] Hoeffding, W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 1948, vol. 19, s. 293 – 325.

[24] Hubert, M., van der Veeken, S. Outlier detection for skewed data. *Journal of Chemometrics*, 2008, vol. 22, no. 3-4, s. 235 – 246.

[25] Hubert, M., van der Veeken, S. Fast and robust classifiers adjusted for skewness. In *COMPSTAT 2010: proceedings in computational statistics: 19th symposium held in Paris, France*. Y. Lechevallier, G. Saporta, eds. Heidelberg: Springer, 2010. s. 1135 – 1142.

[26] Hugg, J., et al. An experimental study of old and new depth measures. In *Proceedings of the eighth Workshop on Algorithm Engineering and Experiments and the third Workshop on Analytic Algorithmics and Combinatorics*. R. Raman, R. Sedgewick, M. F. Stallmann, eds. Miami, FL: Society for Industrial and Applied Mathematics, 2006. s. 51 – 64.

[27] Johnson, D. S., Preparata, F. P. The densest hemisphere problem. *Theoretical Computer Science*, 1978, vol. 6, s. 93 – 107.

[28] Jörnsten, R. Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis*, 2004, vol. 90, s. 67 – 89.

[29] Mosler, K., Hoberg, R. Data analysis and classification with the zonoid depth. In *Data depth: robust multivariate analysis, computational geometry and applications*. R. Y. Liu, R. Serfling, and D. L. Souvaine, eds. 1st edition. New York: American Mathematical Society, 2006. DIMACS series in discrete mathematics and theoretical computer science; 72. s. 49 – 59.

[30] Koshevoy, G. A. The Tukey depth characterizes the atomic measure. *Journal of Multivariate Analysis*, 2002, vol. 83, s 360 – 364.

[31] Koshevoy, G. A., Mosler, K. Zonoid trimming for multivariate distributions. *Annals of Statistics*, 1997, vol. 25, no. 5, s. 1998 – 2017.

[32] Koshevoy, G. A., Mosler, K. Lift zonoids, random convex hulls and the variability of random vectors. *Bernoulli*, 1998, vol. 4, no. 3, s. 377 – 399.

[33] Kosiorowski, D. Robust classification and clustering based on the projection depth function. In *COMPSTAT 2008: proceedings in computational statistics: 18th symposium held in Porto, Portugal* P. Brito, editor. [CD-ROM]. Heidelberg: Physica, 2008. s. 209 – 216.

[34] Li, J., Cuesta-Albertos, J. A., Liu, R. DD-classifier: nonparametric classification procedure based on DD-plot. *unpublished script* [online]. [cited 2011-07-27], available at <http://personales.unican.es/cuestaj/DDPlot_Classification.pdf>.

[35] Liu, R. Y. On a notion of simplicial depth. *Proceedings of the National Academy of Sciences of the USA*, 1988, vol. 85, s. 1732 – 1734.

[36] Liu, R. Y. On a notion of data depth based on random simplices. *Annals of Statistics*, 1990, vol. 18, no. 1, s. 405 – 414.

[37] Liu, R. Y. Control charts for multivariate processes. *Journal of the American Statistical Association*, 1995, vol. 90, no. 432, s. 1380 – 1387.

[38] Liu, R. Y., Parelius, J. M.,Singh, K. Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion). *Annals of Statistics*, 1999, vol. 27, s. 783 – 858.

[39] Liu, R. Y., Singh, K. Rank tests for multivariate scale difference based on data depth. In *Data depth: robust multivariate analysis, computational geometry and applications*. R. Y. Liu, R. Serfling, and D. L. Souvaine, eds. 1st edition. New York: American Mathematical Society, 2006. DIMACS series in discrete mathematics and theoretical computer science; 72. s. 17 – 34.

[40] Massé, J.–C. Asymptotics for the Tukey depth process, with an application to a multivariate trimmed mean. *Bernoulli*, 2004, vol. 10, no. 3, s. 397 — 419.

[41] Massé J.–C., Plante, J.–F. *Package* `depth` *documantation* [online]. Available at <http://cran.r-project.org/>

[42] Mosler, K., Hoberg, R. Data analysis and classification with the zonoid depth. In *Data depth: robust multivariate analysis, computational geometry and applications*. R. Y. Liu, R. Serfling, and D. L. Souvaine, eds. 1st edition. New York: American Mathematical Society, 2006. DIMACS series in discrete mathematics and theoretical computer science; 72. s. 49 – 59.

[43] Romito, F. Elliptically symmetric distributions: a review of achieved results and open issues. In *New developments in classification and data analysis: proceedings of the meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Bologna, September 22-24, 2003*. M. Vichi, et al., eds. Springer, 2005. s. 359 – 366.

[44] Rousseeuw, P. J., Hubert, M. Regression depth (with discussion). *Journal of American Statistical Association*, 1999, vol. 94, s. 388 – 433.

[45] Rousseeuw, P. J., Ruts, I. Algorithm AS 307: bivariate location depth. *Journal of the Royal Statistical Society, Series C*, 1996, vol. 45, no. 4, s. 516 – 526.

[46] Rousseeuw, P. J., Ruts, I., Tukey, J. The bagplot: a bivariate boxplot. *The American Statistician*, 1999, vol. 53, no. 4, s. 382 – 387.

[47] Rousseeuw, P. J.,Struyf, A. Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 1998, vol. 8, s. 193 – 203.

[48] Scott, D. W. *Multivariate density estimation: theory, practice, and visualization*. 1st edition. New York: Wiley, 1992. 317 s. Wiley series in probability and mathematical statistics. ISBN 0-471-54770-0.

[49] Serfling, R. A depth function and a scale curve based on spatial quantiles. In: *Statistical Data Analysis Based On the L_1–norm and Related Methods*. Y. Dodge, editor. Basel: Birkhäuser, 2002. s. 25 – 38

[50] Serfling, R. Depth functions in nonparametric multivariate inference. In *Data depth: robust multivariate analysis, computational geometry and applications*. R. Y. Liu, R. Serfling, and D. L. Souvaine, eds. 1st edition. New York: American Mathematical Society, 2006. DIMACS series in discrete mathematics and theoretical computer science; 72. s. 1 – 16.

[51] Serfling, R. Multivariate symmetry and asymmetry. In *Encyclopedia of statistical sciences*. S. Kotz, N. Balakrishnan, C. B. Read, B. Vidakovic, eds. 2nd edition. New York: Wiley, 2006. s. 5338 – 5345.

[52] Tukey, J. Mathematics and picturing data. In *The Proceedings of the International Congress of Mathematicians held in Vancouver, August 21—29, 1974*. R. D. James, editor. Canadian Mathematical Congress, 1975, vol. 2, s. 523 – 531.

[53] van de Geer, S. A. *Applications of empirical process theory*. 1st edition. Cambridge: Cambridge University Press, 2000. 286 s. Cambridge series in statistical and probabilistic mathematics; 6. ISBN 0-521-65002-X.

[54] Vardi, Y. and Zhang, C.–H. The multivariate L_1–median and associated data depth. *Proceedings of the National Academy of Sciences of the USA*, 2000, vol. 97, no. 4, s. 1423 – 1426.

[55] Vencálek, O. Weighted data depth and its properties. In *Robust 2008*. J. Antoch, G. Dohnal, eds. Praha: Jednota českých matematiků a fyziků, 2009. s. 487 – 494.

[56] Zuo, Y., Serfling, R. General notion of statistical depth function. *Annals of Statistics*, 2000, vol. 28, s. 461 – 482.

[57] Zuo, Y., Serfling, R. On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. *Journal of Statistical Planning and Inference*, 2000, vol. 84, s. 55 – 79.