Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science

Ordering multivariate observations by data depth

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree

BACHELOR OF SCIENCE
in
APPLIED MATHEMATICS

by

MICHEL DIJKSHOORN

Delft, the Netherlands
July 2019

Abstract

In this report the theory of depth functions was researched, as well as an application on this by simulating with multivariate distributions and an application on real-world data about the weather in the Netherlands.

Depth functions are functions that measure data depth and order multivariate observations. They desire properties such as affine invariance, maximality at the center, monotonicity relative to the deepest point and vanishing at infinity. A consequence of this properties is that if a distribution is angularly symmetric about some point μ, then this point is considered as a multivariate median. There are two important depth functions: the halfspace depth and the simplicial depth. The halfspace depth of a point is defined as the smallest probability for which a closed halfspace contains that point. The range of this halfspace depth is $[\frac{1}{n+1}, \frac{1}{2}]$, where n is the number of observations. For the simplicial depth, a distinction is made in the sample and the population version. The sample simplical depth of a point is defined as the number of closed simplices in which that point is contained, divided by the total number of simplices. The population version is defined analogously, however with a probability instead. Its maximum is 2^{-d}, with dimension d. As the number of observations tends to infinity, the sample simplicial depth tends to the population simplicial depth. Furthermore, contours for which the depth is constant provide a good visualization of the depth, provided that the dimension allows visualization, e.g. $d \leq 3$. These contours are nested within each other. For elliptical distributions the shapes of halfspace contours and isodensities coincide, as they are both ellipses. Finally, the outlyingness is defined as the inverse of the depth minus 1, such that the ordering is reversed.

This theory is applied by simulating with multivariate distributions. The first distribution to be considered is the bivariate normal distribution, in which the distinction of independent and dependent variables is made. For both situations, the points with the lowest depth are situated far away from the mean, in fact they are almost equally distanced from it. They highest depth points are near the mean. The contours and isodensities are indeed ellipses. The halfspace contours coincide with the isodensities. This is true for the simplicial depth as well, though the outer contour is not shaped the same. Particularly, for the independent variables the ellipses are circles. Another distribution is the bivariate Student's t-distribution with independent variables. The isodensities are shapes as squares and the halfspace depth contours are shaped as stars with four ends. The last distribution being considered is the normal distribution in the fifth dimension. Despite not being able to visualize these observations, nevermind the depth, we can however compare the depth with the density. After a modification of the depth, it turns out they have the same values.

The final part of this report is about the application of the theory to data of the weather. The first data set to be considered is the maximal temperatures for every day in July for the years 1901–2018. The month July is cut into smaller parts, such that we have 25 intervals of 7 days. For all years, the depth can be computed for all 25 intervals. Since the depth range is very dense, the outlyingness is used. 2018 turns out not as outlying as thought before: only for the last week the outlyingness is high. More years are considered and their depths corresponds somewhat to their mean temperatures. The second data set to be considered is a bivariate data set: the highest average wind speed in an hour and the highest amount of precipiation in an hour for every day in the years 1906-2018. When both of these are high at the same, it is considered as severe weather. The number of observations is lowered in order to compute contour lines. These contours showed a good notion of the distribution of all data.

Preface

This report is written for the completion of the Bachelor Applied Mathematics at the Delft University of Technology. It was primarily written in May and June 2019. The reason I chose the subject of depth statistics is because I could apply it to weather data, which I have always found very interesting. The main goal of this thesis is for the author to understand the theory, rather than being a source for the public. Nevertheless, this report may be read by anyone who has knowledge of Bachelor level statistics or higher. Finally, I would like to thank my supervisor Juan-Juan Cai for helping me throughout the whole project.

Contents

1 Introduction

The month July of 2018 broke several records in the Netherlands [1]. According to measurements of the KNMI (Royal Netherlands Meteorological Institute), it was a record breaking month. The average temperature was 20.7 °C against an overall average of 17.9 °C. Yet, this was not the hottest July ever: 2006 and 1994 tapped 22.3 °C and 21.4 °C respectively. It is still exceeding, because the KNMI has been measuring temperatures since 1901, so the year of 2018 has the third highest average July temperature out of 118 measured years. Moreover, on July 26th the highest temperature ever was measured in De Bilt, with 35.7 °C. Another record is the precipitation record: a poor 5.5 mm of percipitation was measured over the whole month, instead of a normal 81.1 mm: also a new record. Finally, it was the sunniest July ever, with 341 hours of sun, compared to an average of 206 hours.

All these records brought big problems: a great risk on wildfires, harvest failures, dying plants etc. [2]. It could be useful to prepare for such months, so one may ask themselves: how often do such extreme months occur? To answer this question, the theory of depth functions can be used. The theory of depth functions is used to order multivariate observations by data depth. Depth is used to measure centrality. If data has high depth, then it is very central. In fact, an observation with the highest depth is considered to be a multivariate median. It also means that extreme data can be measured by low depth. A depth function orders this data. There are some properties that these depth function should suffice, but there exists a large amount of depth functions. Two important ones are introduced by Tukey in 1975 [3] and Liu in 1990 [4].

In Section 2 the theory of depth functions is explained. Two specific depth functions, the halfspace and simplicial depth function, will be researched here. There are a couple of theorems and lemmas in this section and the proofs of these are either found in Appendix A or referenced. In Section 3 the theory is applied to simulations with multivariate distributions, e.g. the bivariate normal distribution. In Section 4, some real-world data is looked into and the theory is again applied on this, to eventually say something about the frequency of extreme weather, particularly of temperatures and severe weather which in this case is heavy wind combined with heavy precipitation.

2 Depth functions

The main goal of this project is to find a way to measure the data depth in multivariate data. The data depth is a measure of how deep a data point (or an observation) lies in a data cloud. For instance one may consider the univariate normal distribution. If a data point is located near the median, then we say that this point is deep in data, since most of the observations lie around the median. Observations far away from the median are not deep in the data.

To find out how deep a point lies in a data cloud, the data set needs to be ordered. In the univariate case, i.e. observations in $\mathbb{R}^1$, one can order the data in value from small to large. With bivariate or multivariate data, that will not work anymore, since each point consists of more than one value. Thus, another method is required to order this kind of data. This is done by introducing a depth function. There exist several kinds of depth functions [5], however two of them will be discussed below: the halfspace depth and the simplicial depth, in which we make the distinction between the sample and the theorical one.

2.1 Key properties of a depth function

It is just stated that there exist multiple depth functions. These depth functions should all contain certain properties to satisfy some sort of ordering of the data. A depth function is denoted as $D(x|F_X)$, with $x \in \mathbb{R}^d$ and where X is a random vector in $\mathbb{R}^d$, with cdf F. The properties, as proposed in [4], are stated below:

P1. *Affine invariance.* $D(Ax + b|F_{AX+b}) = D(x|F_X)$, for any nonsingular matrix A and vector b. It means that any depth function is invariant under affine transformations, e.g. translations, rotations etc.

P2. *Maximality at the center.* If μ is the center (point of symmetry) of F, then $D(\mu|F) = \sup_{x\in\mathbb{R}^d} D(x|F)$.

P3. *Monotinicity relative to the deepest point.* If μ is the point of maximal depth of F, then $D(x|F) \leq D(\mu + \alpha(x-\mu)|F)$, with $\alpha \in [0,1]$. Note that $\|\mu - (\mu + \alpha(x-\mu))\| = |\alpha| \cdot \|x-\mu\| \leq \|x-\mu\|$. This means that points closer to the deepest point have higher depth.

P4. *Vanishing at infinity.* $D(x|F) \to 0$ as $\|x\| \to \infty$.

In P2, the symmetry is not clearly defined, because there are different ways to define multivariate symmetry. Here, two will be considered, the angular symmetry and the halfspace symmetry.

A random variable X in $\mathbb{R}^d$, or its distribution F, is called angular symmetric about $\mu \in \mathbb{R}^d$ if and only if the random variables $(X-\mu)/\|X-\mu\|$ and $-(X-\mu)/\|X-\mu\|$ are equally distributed. X is defined to be halfspace symmetric if $P(X \in H) \geq \frac{1}{2}$ for every closed halfspace H containing b. If F is angularly symmetric about μ, then any hyperplane passing through μ divides $\mathbf{R}^d$ into two halfspaces with equal probability, $\frac{1}{2}$. Thus, the center of angular symmetry is some sort of multivariate median. Moreover, angular symmetry implies halfspace symmetry [6].

2.2 Halfspace depth

The first depth function being discussed will be the halfspace depth. It is also known as the Tukey depth, since it was introduced by Tukey in 1975 [3]. This method is explained by first giving an intuitive idea and a formilization follows afterwards.

The halfspace depth is based on the centrality of data. Consider a data cloud in $\mathbb{R}^d$. The median of the data cloud should be in the absolute center of that cloud. Consider a symmetric model. As a consequence of the symmetry of the data cloud, every hyperplane (lines in $\mathbb{R}^2$, regular planes in $\mathbb{R}^3$ etc.) that goes through the median should be cutting the data cloud

exactly in half. This hyperplane cuts $\mathbb{R}^d$ into two spaces, which are called halfspaces. If such a hyperplane has the median on its boundary, then both halfspaces should contain the same amount of data, theoretically. This is not the case for observations other than the median, where some halfspaces that have such observations on its boundary contain more than half of the data and some less than a half. What follows is that when the amount of observations in halfspaces differ more, then every observation on the boundary are considered more outlying. The idea of the halfspace depth is that for an observation, every halfspace that has this observation on its boundary is researched and the one containing the least data will be used.

As a formalization, the halfspace depth of a point z is defined to be the smallest probability of a halfspace containing z:

$$D_H(z) = \inf\{P(H)|\ H \text{ is a closed halfspace, } z \in H\} \tag{1}$$

Note that z does not necessarily have to be on the boundary of a halfspace, according to the definition. However, if H is a closed halfspace containing z, but z is not on the boundary of H, and if H minimizes the amount of observations contained, then this halfspace can be restricted by moving its boundary parallel towards z. The probability will remain the same and now z is on the boundary of the new halfspace. Thus, without loss of generality, it may be assumed that the boundary of a closed halfspace H which contains z and minimizes $P(H)$ passes through that point z.

This depth function can be expressed in terms of its distribution function for the univariate case. A univariate halfspace is a half line, i.e. a straight line extending from a point indefinitely in one direction only. Consider the following theorem:

Theorem 1. *If F is a distribution in $\mathbb{R}^1$, then*

$$D_H(z) = \min\{F(z), 1 - F(z)\} \in \left[0, \frac{1}{2}\right].$$

Moreover, the depth is maximized in the median μ with maximum value $\frac{1}{2}$.

The proof may be found in appendix A. This theorem may be extended to the multivariate case.

It is also possible to give an upper bound for the half space depth. In fact, a maximum may be given, under a couple of conditions. Firstly, an upper bound will be given in the following lemma:

Lemma 1. *For any probability P with a density we have $D_H(z) \leq \frac{1}{2}$ for any $z \in \mathbb{R}^d$.*

The proof may be found in appendix A. Furthermore, the following theorem provides a maximum halfspace depth:

Theorem 2. *If P has a density and is angularly symmetric about μ, then*

$$\max_z D_H(z) = D_H(\mu) = \frac{1}{2}.$$

The proof may be found in appendix A. Note that Theorem 1 is in line with both Lemma 1 and Theorem 2. Furthermore, it is stated in [7] that any probability P without further assumptions (e.g. does not need a density) has a lower bound for the maximum depth: $\frac{1}{n+1}$, where n is the number of observations.

2.3 Sample simplicial depth

The sample simplicial depth is a way to measure depth based on simplices. It may be used for data which has an unknown distribution function. The simplicial depth was proposed by Liu in 1990 [4].

2.3.1 Bivariate observations

To explain this method, the bivariate case is considered. Let $X_1, \ldots, X_n$ be a bivariate data set, i.e. $X_i \in \mathbb{R}^2$ for all $1 \leq i \leq n$. Graphically, this is a two-dimensional field of points. For all $1 \leq i, j, k \leq n$, where i, j and k are unequal, a triangle can be created. The total number of these triangles is equal to the amount of combinations of three out of a set of n elements. Thus, in total there are $\binom{n}{3}$ such triangles. In such a triangle, there may or may not be other observations contained. For every point $x \in \mathbb{R}^2$ the number of triangles that contain x can be determined. The event that a triangle formed out of the points X_i, X_j and X_k contains x is denoted by $x \in \Delta(X_i, X_j, X_k)$. Now, let I be the indicator function, i.e. $I(A) = 1$ if event A occurs and $I(A) = 0$ otherwise. The total number of triangles that contain x may now be expressed as

$$\sum_{1 \leq i,j,k \leq n} I(x \in \Delta(X_i, X_j, X_k))$$

In sake of completeness, we say that if an observation is an angle of a triangle, then it is contained in that triangle, in other words, all triangles are closed. The number of triangles that contain a point x can be scaled into $[0, 1]$ by dividing it by the total number of triangles. We then get

$$\widehat{D}_S(x|X_1, \ldots, X_n) = \binom{n}{3}^{-1} \sum_{1 \leq i,j,k \leq n} I(x \in \Delta(X_i, X_j, X_k)) \tag{2}$$

This is called the sample simplicial depth. Intuitively, it is a measure of the deepness of an observation in a data set. That is because when an observation is contained in many triangles, it is surrounded by many other observations, so it is near the median and hence it lies deep in the data.

2.3.2 Multivariate observations

The bivariate case has been considered because the concept may then be visualized easily. The method works analogously for multivariate cases (i.e. in $\mathbb{R}^d$ with $d \geq 2$).

Let $X_1, \ldots, X_n$ be a multivariate data set in $\mathbb{R}^d$. For $d + 1$ observations a simplex can be created with these observations as vertices. This simplex is indicated as $S(X_{i_1}, \ldots, X_{i_{d+1}})$. In the bivariate case ($d = 2$), the simplex was a triangle. For $d = 3$, the simplex is a tetrahedron. The total number of simplices that can be created with n observations is $\binom{n}{d+1}$. The scaled number of simplices that contain $x \in \mathbb{R}^d$ may now be expressed as

$$\widehat{D}_S(x|X_1, \ldots, X_n) = \binom{n}{d+1}^{-1} \sum_{1 \leq i_1 < \ldots < i_{d+1} \leq n} I(x \in S(X_{i_1}, \ldots, X_{i_{d+1}}))$$

When implementing this function, a problem will arise. How does one verify whether a point $x \in \mathbb{R}^d$ is contained in a simplex $S(X_1, \ldots, X_{d+1})$? This can be obviously observed by the human eye, however with an implementation it is a little harder.

This is a problem in the geometry and may be solved through linear algebra. Consider the following system of linear equations:

$$\begin{cases} x = \alpha_1 X_1 + \cdots + \alpha_{d+1} X_{d+1}; \\ \alpha_1 + \cdots + \alpha_{d+1} = 1. \end{cases}$$

Solving this system of $d+1$ equations with $d+1$ unknowns (the alphas) provides a unique solution if the simplex is nondegenerate [4]. Then the following theorem holds: x is contained in the simplex if and only if all $\alpha_1, \ldots, \alpha_{d+1}$ are positive.

2.4 Population simplicial depth

For the theoretical simplicial depth, it is assumed that the data has a known distribution function. Note that $\widehat{D}_S(x)$ as in the sample simplicial depth is the depth based on a sample. It is in fact the empirical version of the probability

$$D_S(x|F) = P_F(x \in S(X_1, \ldots, X_{d+1})),$$

where $X_1, \ldots, X_{d+1}$ are i.i.d. with continuous cdf F. $D_S(x)$ is simply called the simplicial depth.

It is desired to express the depth in terms of cdf F. Only the univariate case will be treated here, by the following theorem:

Theorem 3. *Let $X_1, X_2 \in \mathbb{R}^1$ be i.i.d. with univariate continuous cdf F and let $\overline{X_1 X_2}$ be the line segment connecting X_1 and X_2. The simplicial depth will then be*

$$D_S(x|F) = P_F(x \in \overline{X_1 X_2}) = 2F(x)(1 - F(x)).$$

Moreover, the depth is maximized in the median with maximum value $\frac{1}{2}$.

The proof may be found in appendix A. For the multivariate case, the depth won't be expressed in terms of F. Instead, other properties and consequences of the multivariate simplicial depth will be treated.

Theorem 4. *If F is absolutely continuous and angularly symmetric about the origin, then $D_S(\alpha x|F)$ is a monotone nonincreasing function in $\alpha \geq 0$ for all $x \in \mathbb{R}^d$.*

The proof is given in [4]. Note that because of property P1, as stated in Section 2.1, the above theorem may be generelized about any point in $\mathbb{R}^d$, instead of just the origin. The next theorem provides a maximum of the simplicial depth for any Euclidean space:

Theorem 5. *If F is an absolutely continuous distribution on $\mathbb{R}^d$ and it is angularly symmetric about $\mu \in \mathbb{R}^d$, then $D_S(\mu|F) = 2^{-d}$.*

The proof is given in [4]. From both Theorem 4 and 5, it follows that if F is an angularly symmetric (about μ) distribution, then the maximum is attained at μ with $D(\mu) = 2^{-d}$, hence for any $z \in \mathbb{R}^d$ we have $D(z) \leq 2^{-d}$. Moreover, Theorem 3 is in line with this property.

Both the sample and the theoretical simplicial depth have been considered. However, they lack a connection at this point. Therefore, another theorem is considered: the uniform consistency of $\widehat{D}_S(x)$.

Theorem 6. *If $X_1, \ldots, X_n$ are i.i.d. with absolutely continuous cdf F and if its density f is bounded, then*

$$\sup_{x \in \mathbf{R}^d} |\widehat{D}_S(x|X_1, \ldots, X_n) - D_S(x|F)| \to 0 \qquad \text{a.s.} \qquad \text{as } n \to \infty.$$

The proof is given in [4]. Note that a.s. stands for almost surely.

2.5 Contours

So far, two depth functions have been discussed, including some of their properties. It is clear from the properties stated in section 2.1 that these depth functions provide a kind of center-outward ordering of data. This may be visualized through contours.

Consider a bivariate data cloud. We are interested in visualizing the data depth of the data cloud. One way to do this is to associate every observation with its depth. Observations with the same depth may be connected with each other. In that way, several curves around the highest depth observation will arise. These curves are called contour lines or simply contours, since they join every point with the same depth value. This may be extended and formalized to the multivariate case.

Given a distribution F in $\mathbb{R}^d$, a point $x \in \mathbb{R}$ and some depth function $D(x|F)$. For $\alpha > 0$, the boundary of the set $\{x \in \mathbb{R}^d |\ D(x|F) \geq \alpha\}$ defines a contour for which all observations have depth α. The region

$$D^\alpha(F) := \{x \in \mathbb{R}^d |\ D(x|F) \geq \alpha\}$$

is defined [6] as the α-trimmed region and its boundary

$$\partial D^\alpha(F) := \{x \in \mathbb{R}^d |\ D(x|F) = \alpha\}$$

as the α-contour. It is analogously defined for sample depth functions.

All trimmed regions and contours are related to each other. From Theorem 4 it follows that $D^\alpha(F) \subset D^\beta(F)$ if $\alpha \geq \beta > 0$ and hence the contours are nested within eachother. They move further and further away from the center as α decreases. Moreover, from this it follows that every $D^\alpha(F)$ is connected.

The contours can be made for density functions instead of depth functions as well. Then they consist of every point with the same density value, so the density contours are curves with a constant density value. These density contours are called isodensity lines or isodensanes. For special distributions such as the elliptical distribution, depth contours posess the same shape as the isodensity lines. This is proved in Theorem 7.

Firstly, elliptical distributions are defined. A distribution F in $\mathbb{R}^d$ is elliptically distributed with location-scatter parameter (μ, Σ) if its density f is of the form

$$f(x) = c|\Sigma|^{-\frac{1}{2}} h((x-\mu)^T \Sigma^{-1}(x-\mu)).$$

where $\mu \in \mathbb{R}^d$, Σ is a positive definite $p \times p-$matrix, c is a constant and h is some univariate function. It is denoted as $F \sim E_d(h; \mu, \Sigma)$. The isodensity lines for these distributions are ellipses [8].

Theorem 7. *Suppose that $F \sim E_d(h; \mu, \Sigma)$ and that $D(x|F)$ is a depth function for which properties P1 and P2 (with maximum value μ) hold. Then:*

1. *$D(x|F)$ is of form*

$$D(x|F) = h((x-\mu)^T \Sigma^{-1}(x-\mu))$$

for some nonincreasing function h, and $D^\alpha(F)$ is of form

$$D^\alpha(F) = \{x \in \mathbb{R}^d |\ (x-\mu)^T \Sigma^{-1}(x-\mu) \leq r_\alpha^2\}$$

for some r_α.

2. *$D(x|F)$ is strictly decreasing on any ray originating from the center if and only if*

$$\{x \in \mathbb{R}^d |\ D(x|F) = \alpha\} = \{x \in \mathbb{R}^d |\ (x-\mu)^T \Sigma^{-1}(x-\mu) = r_\alpha^2\}$$

3. *If, moreover, F is normally distributed and D denotes the halfspace depth, then* $r_\alpha = \Phi^{-1}(1-\alpha)$.

The proof is given in [6]. From Theorem 7, it follows that for elliptical distributions the depth contours coincide with the isodensity lines. Therefore, the shape of depth contours provide information about the density of the observations.

So far, we have seen that depth functions provide a good way to order multivariate data. Property P3. from section 2.1 makes sure that observations have lower depth when moving away from the median and that is exactly what we desire from depth functions.

Furthermore, contours are very useful for visualization of the depth in a data cloud. With these contours it is possible to create a region in which all data has a minimal depth and therefore a region with all data which has a maximal depth as well. In this way, outliers can be classified.

Firstly, the data are formally ordered. Let $X_1 \ldots, X_n$ be observations in $\mathbf{R}^d$ and let $X_{\lfloor 1 \rfloor}, \ldots, X_{\lfloor n \rfloor}$ be the same data, but ordered from high depth to low depth (with $X_{\lfloor 1 \rfloor}$ as the highest and $X_{\lfloor n \rfloor}$ as the lowest depth). Outliers are ordered exactly the other way around. An outlyingness function is proposed in [9], defined as follows:

$$Out(z|F) = \frac{1}{D(z|F)} - 1.$$

Points outside the region $D^\alpha(F)$ having outlyingness greater than $\frac{1}{\alpha} - 1$ can be regarded as outliers of a specified level α. For the halfspace depth, we have that $Out_H : \mathbb{R}^d \to [1, \infty)$. For the simplicial depth, we have that $Out_S : \mathbb{R}^d \to [2^d - 1, \infty)$. Since the range differs for every depth function, it is not possible to compare the outlyingness for different depth functions. Of course, the ordering of a data set for different depth functions can be compared.

In sake of visualization, an example is used [9]. Consider a bivariate data set with the variables governmental debt as a percentage of the GDP (general domestic product) and unemployment rate of EU-27 countries in 2011. In figure 1 one can see these variables plotted against each other. The contours are given as well, and with these one can easily verify that there are 5 outliers with respect to the outer drawn contour: Estonia, Luxembourg, Austria, Spain and Greece.

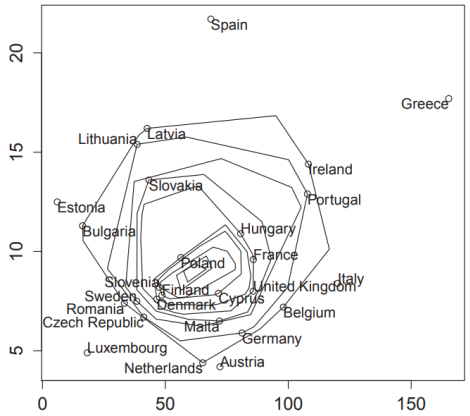

Figure 1: Governmental debt (in % of GDP) on the x-axis vs. the unemployment rate (in %) on the y-axis. Some contours of the halfspace depth are drawn [9].

3 Simulation with multivariate distributions

In this section the theory from Section 2 is applied by simulating, i.e. generating a lot of random samples from some distribution. The idea is to compute simulations from a particular multivariate distribution and verify the results found in Section 2 for both depth functions. Simulations will be performed in the statistical programming language R.

3.1 Bivariate Normal distribution

The first distribution being considered is the multivariate normal distribution. It is defined [10] as follows: A random vector $\mathbf{X} = [X_1, \ldots, X_d]^T \in \mathbb{R}^d$ is said to have a multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ if its probability density function is given by:

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

for some $\mathbf{x} \in \mathbb{R}^d$. It is denoted as $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$. The covariance matrix Σ should be a positive definite $d \times d$-matrix (i.e., $\Sigma = \Sigma^T$ and $\mathbf{x}^T\Sigma\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^d$ such that $\mathbf{x} \neq \mathbf{0}$). If we just focus on the bivariate case, let $\mathbf{X} = [X, Y]^T$ and

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

If Σ is a diagonal matrix, then X and Y are uncorrelated and therefore independent [10].

3.1.1 Independent variables

The first example will be a bivariate normal distribution with independent variables. Let

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Note that Σ_1 is the identity matrix and is therefore positive definite. Firstly, $n = 1000$ random samples are drawn from $\mathbf{X} = [X, Y]^T \sim \mathcal{N}_2(\boldsymbol{\mu}_1, \Sigma_1)$. Every paired sample is denoted as $z_i = (x_i, y_i)$, where $i \in \{1, \ldots, n\}$. Their ranges are $x_i \in [-2.005; 4.231]$ and $y_i \in [-1.936; 4.001]$. For every pair, both the halfspace and simplicial depth may be computed (denoted respectively as $D_H(z_i)$ and $D_S(z_i)$), as defined in equations 1 and 2. See table 4 in appendix B for both depth values for specific samples. These include all samples with the lowest halfspace depth and the 10 samples with the highest halfspace depth.

In figure 2, all samples are plotted. Moreover, all samples from table 4 are highlighted with either red colors (all smallest halfspace depth values) or green colors (10 largest halfspace depth values). The blue lines represent the means of the variables and they cross each other in the center. One can see that the red samples are far away from the center and that the green samples are near the center, exactly what would be expected.

The halfspace depth values range from 0.001 to 0.481. This is in line with the theory. The smallest halfspace depth value should be at least $1/n = 0.001$, because there could be a halfspace which contains no points other than the point itself. There are 10 such points. The highest halfspace depth value should be at most 0.5. However, this will almost never be attained when sampling. Instead, the highest halfspace depth value is a tiny bit smaller. The simplicial depth values range from 0.003 to 0.253. This is not completely in line with the theory. The smallest simplicial depth value should be at least the number of simplices that contain some

point divided by the total number of simplices, i.e. $\frac{1}{2}(n-1)(n-2)\cdot\binom{n}{3}^{-1} = \frac{3}{n} = 0.003$. So the smallest value is correct. However, the largest value exceeds the upper bound given in Theorem 5, which is $2^{-2} = 0.25$. This will be discussed in Section 6. The smallest depth values match both depth functions. However, for larger depths not all depth values match anymore, but they differ not much either. For instance, the fourth highest halfspace depth corresponds to the second highest simplicial depth. In conclusion, both depth functions provide a different kind of ordering.

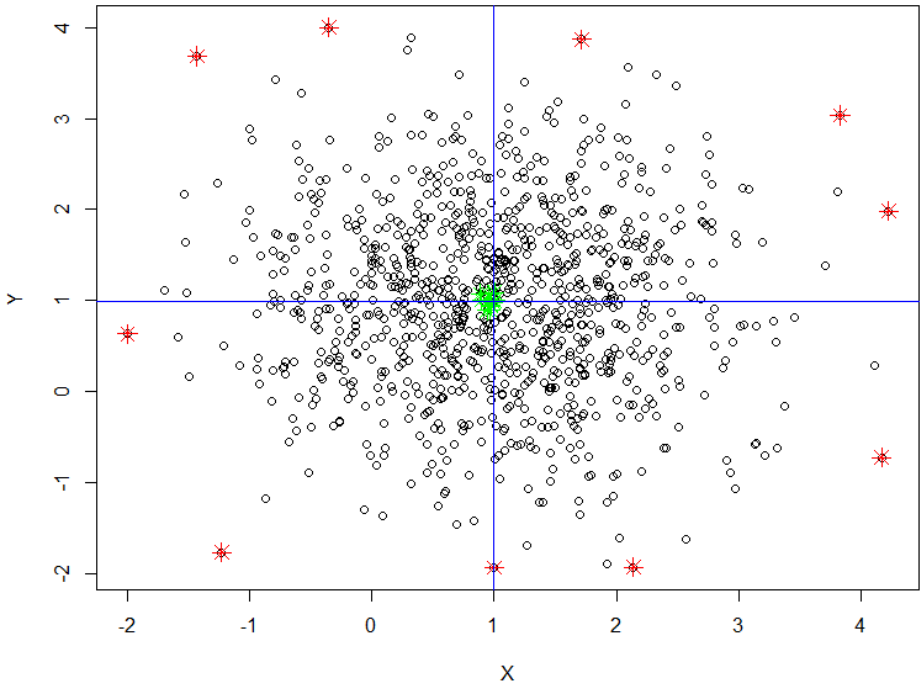

Figure 2: 1000 random samples from $\mathcal{N}_2(\boldsymbol{\mu}_1, \Sigma_1)$, with all smallest halfspace depth values in red and the 10 largest halfspace depth values in green.

The red points in figure 2 are somewhat far distanced from the center. All of these are more or less connected by a circle. Since these points have the same depth, they should indeed all be on an ellipse (or specifically a circle [6]). Generally, every halfspace depth contour should coincide with isodensities according to Theorem 7, and the following equality holds for the halfspace depth:

$$\{\mathbf{x} \in \mathbb{R}^d |\ D_H(\mathbf{x}|F) = \alpha\} = \{\mathbf{x} \in \mathbb{R}^d |\ (\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) = (\Phi^{-1}(1-\alpha))^2\}$$

In order to match a contour of level α with an isodensity of the same distribution, one must take the following level for an isodensity: $\frac{1}{2\pi}\exp(-\frac{1}{2}(\Phi^{-1}(1-\alpha))^2)$. This follows directly from the definition of the density function. In figure 3 the samples are again plotted, but now with halfspace and simplicial depth contours and isodensities. For both depths, there are five levels of contours plotted: $\alpha \in \frac{1}{n}\{1, 5, 20, 50, 100\}$. Five levels of isodensities are plotted as well, with constant values $\frac{1}{2\pi}\exp(-\frac{1}{2}(\Phi^{-1}(1-\alpha))^2)$. Unfortunately, it is unknown for which values the contours of the simplicial depth and the isodensities coincide (at least unknown to the author). The contours are nevertheless plotted, because the shape of these contours is still interesting.

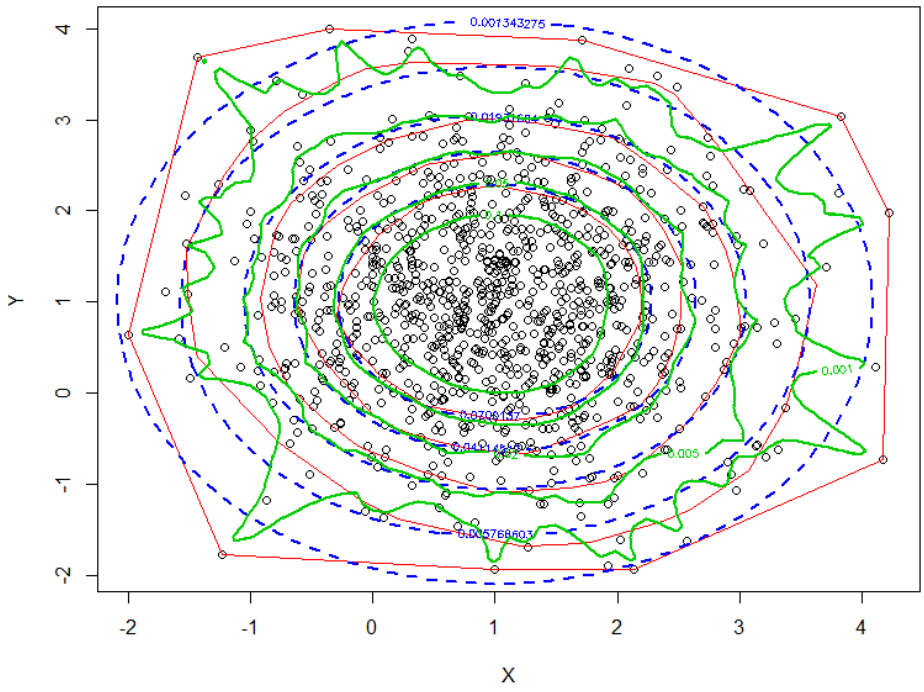

Figure 3: 1000 random samples from $\mathcal{N}_2(\boldsymbol{\mu}_1, \Sigma_1)$, with halfspace depth contours in red solid lines, simplicial depth contours in green solid lines and isodensities in dashed blue lines.

All of the contours (and isodensities as well) are nested within each other. This is in line with Theorem 4. The outer contour of the halfspace depth is the least shaped as a circle. That makes sense, since the least data is located there. As the contour level rises, the contours become more shaped like circles and coincide with the isodensities better. For the simplicial depth contours, only their shape can be compared to the simplicial depth contours and the isodensities. Apart from the outer contour, all contours are more or less shaped as circles again. For the outer contour, one can see that it is shaped as a circle with shoot outs to points with the lowest depth values. Initially, this would just be a circle. Assume that there is a point just outside this circle, call it z_{out}, and let z_{in} be some point on the circle which is very close to z_{out}, z_{in}. The inclusion of z_{out} makes the depth of z_{in} rise, because it creates a lot of triangles in which z_{in} is contained. This is not the case for points on the circle further away from z_{out}, and hence the outer contour contains some shootouts to outliers.

3.1.2 Dependent variables

The second example will be a bivariate normal distribution with dependent variables. Let

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Almost the same $\boldsymbol{\mu}$ and Σ are taken as the previous example, with one exception: X and Y are not uncorrelated anymore. Obviously, $\rho \neq 0$ in order to be dependent, so another ρ must be chosen. However, there should be another condition: Σ_2 should still be positive definite. Let's first compute the eigenvalues of this matrix:

$$\det(\Sigma_2 - \lambda I) = (1-\lambda)^2 - \rho^2 = 0 \quad \Longleftrightarrow \quad \lambda = 1 \pm \rho.$$

It follows that the eigenvalues are positive if and only if $-1 < \rho < 1$. If this holds, then, keeping in mind that Σ_2 is a real symmetric matrix, Σ_2 is positive definite [11]. Thus, any $|\rho| < 1$ may be chosen. Take for example $\rho = 0.5$.

Again, $n = 1000$ random samples are drawn, now from $\mathcal{N}_2(\boldsymbol{\mu}_2, \Sigma_2)$. Every paired sample is denoted again as $z_i = (x_i, y_i)$, where $i \in \{1, \ldots, n\}$. Their ranges are now $x_i \in [-2.511; 4.180]$ and $y_i \in [-2.109; 4.543]$. See table 5 in appendix B for both depth values for the same kind of samples as before.

Figure 4 is generated analogously to Figure 2. The halfspace depth values range from 0.001 to 0.481. This is in line with the theory. The simplicial depth values range from 0.003 to 0.253. This is not completely in line with the theory, for the same reason as before. The smallest depth values still match both depth functions. This is not the case for larger depths, however the difference is not high.

So far, the normal distribution with dependent variables provides the same results as the normal distribution with independent variables. The only difference is that now, the samples are not scattered around the center in circles, but in ellipses.

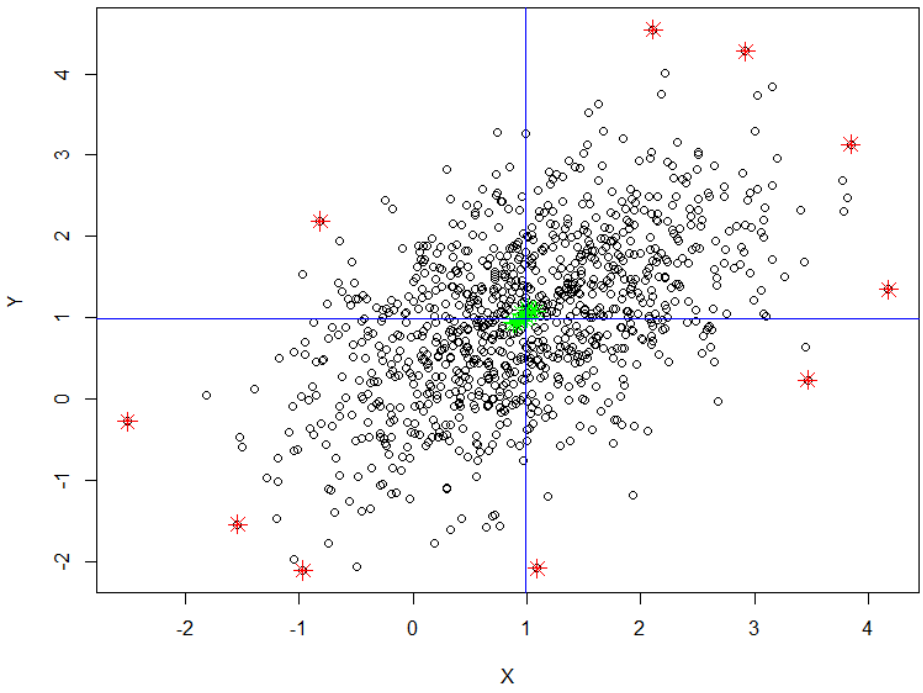

Figure 4: 1000 random samples from $\mathcal{N}_2(\boldsymbol{\mu}_2, \Sigma_2)$, with all smallest halfspace depth values in red and the 10 largest halfspace depth values in green.

The red points in figure 4 are not equally distanced from the center, but there is a pattern. All of these are more or less connected by a ellipse. The same equality holds for the halfspace depth contours and the isodensities. However, different isodensity levels should be taken, because the density function is slightly different. In order to match a contour of level α with an isodensity of the same distribution, one must take the following level for an isodensity:

$\frac{1}{2\pi|\Sigma|^{1/2}} \exp(-\frac{1}{2}(\Phi^{-1}(1-\alpha))^2)$. See figure 5 for all of these contours and isodensities, with the same α levels (for the contours) as before. Also in this case one can only compare the shape of the simplicial depth contours with the other contours/isodensities. The result is very similar as with independent variables, but instead of circles the contours are ellipses. Again, the halfspace depth contours have the same shape as the isodensities. Note however, that they seem to coincide less better than with the independent variables, especially the inner contour. This is discussed in Section 6.

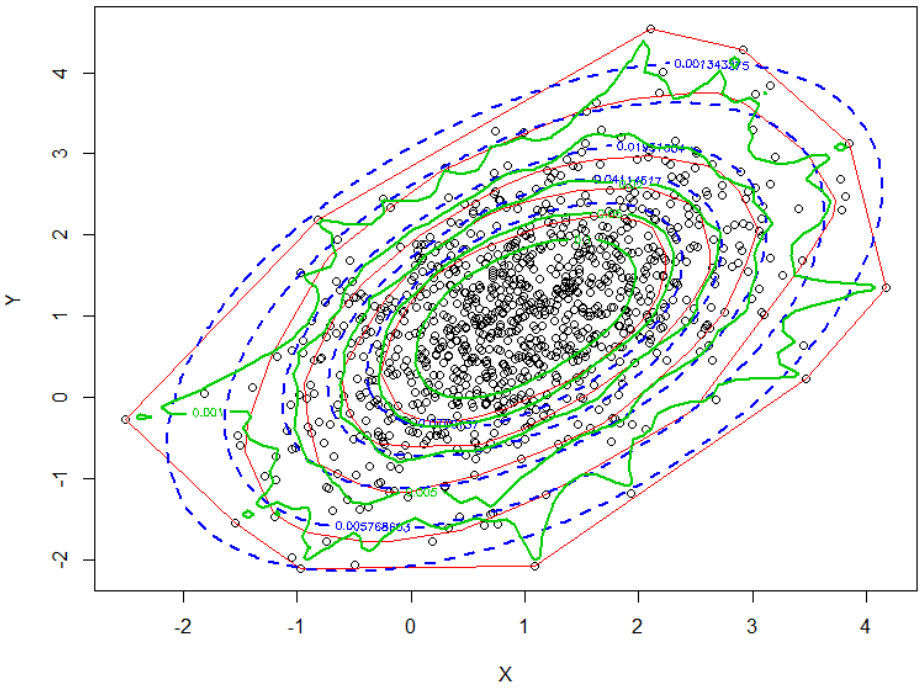

Figure 5: 1000 random samples from $\mathcal{N}_2(\boldsymbol{\mu}_2, \Sigma_2)$, with halfspace depth contours in red solid lines, simplicial depth contours in green solid lines and isodensities in dashed blue lines.

3.2 Bivariate Student's t-distribution

The second distribution being considered is the multivariate Student's t-distribution. It is commonly defined [12] as follows: A random vector $\mathbf{X} = [X_1, \ldots, X_d]^T \in \mathbb{R}^d$ is said to have a multivariate Student's t-distribution with parameters the mean $\boldsymbol{\mu} \in \mathbb{R}^d$, the scale matrix $\Sigma \in \mathbb{R}^{d \times d}$ and the degrees of freedom $\nu \in \mathbb{N}$ if its probability density function is given by:

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma, \nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{d/2}|\Sigma|^{1/2}} \left(1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-(\nu+d)/2}$$

for some $\mathbf{x} \in \mathbb{R}^d$. Note that in general, Σ is not the covariance of $\mathbf{X}$. It is denoted as $\mathbf{X} \sim t_\nu(\boldsymbol{\mu}, \Sigma)$. The Student's t-distribution is related to the normal distribution, because the latter is a special case of the first. If the degrees of freedom ν tends to infinity, the t-distribution converges to a normal distribution [13]. Despite this relation, there are some differences. Firstly, a t-distribution has a different tail behavior than the normal distribution: it has heavy tails [13].

As a consequence, the t-distribution is better to generate outliers. Furthermore, dependency behavior works differently for these distributions. It was already stated in this subsection that whenever Σ is a diagonal matrix, the variables of a normal distribution are uncorrelated and therefore by default independent. This is not the case for arbitrary t-distributions.

Here, we will only consider the bivariate distribution with 1 degree of freedom ($\nu = 1$):

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma, \nu = 1) = \frac{1}{2\pi|\Sigma|^{1/2}} \left(1 + (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)^{-3/2}.$$

Note that $\Gamma\left(\frac{3}{2}\right)/\Gamma\left(\frac{1}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right)/\Gamma\left(\frac{1}{2}\right) = \frac{1}{2}$.

It is desired to look into two independent Student's t-distributions with 1 degree of freedom. The problem is that this pdf is not easily expressed for independent variables. Therefore, another definition of the Student's t-distribution is used. Consider the joint pdf for n independent, zero-mean (location parameter $\boldsymbol{\mu} = \mathbf{0}$) Student's t pdfs with shape parameter $\boldsymbol{\nu}$, as defined in [14]:

$$f(\mathbf{x}|\boldsymbol{\nu}) = \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\Gamma\left(\frac{\nu_i}{2}\right)\sqrt{\pi\nu_i}} \left(1 + \frac{x_i^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}}.$$

For $\boldsymbol{\nu} = \mathbf{1}$, and $d = 2$, we have that

$$f(x, y|\boldsymbol{\nu} = \mathbf{1}) = \frac{(\Gamma(1))^2}{\pi(\Gamma(\frac{1}{2}))^2} \left(1 + x^2\right)^{-1} \left(1 + y^2\right)^{-1} = \frac{1}{\pi^2(1 + x^2)(1 + y^2)},$$

since $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$. In figure 6, some depth contours and isodensities are drawn for this density function. The depth contours are drawn with levels $\alpha \in \frac{1}{n}\{5, 10, 20, 50\}$, where $n = 1000$ is the sample size. In [7], it is stated that the shapes of the isodensities are in fact squares, and the shapes of the halfspace depth contours are shown to be a kind of star with four ends. The shapes of figure 6 are in line with this.

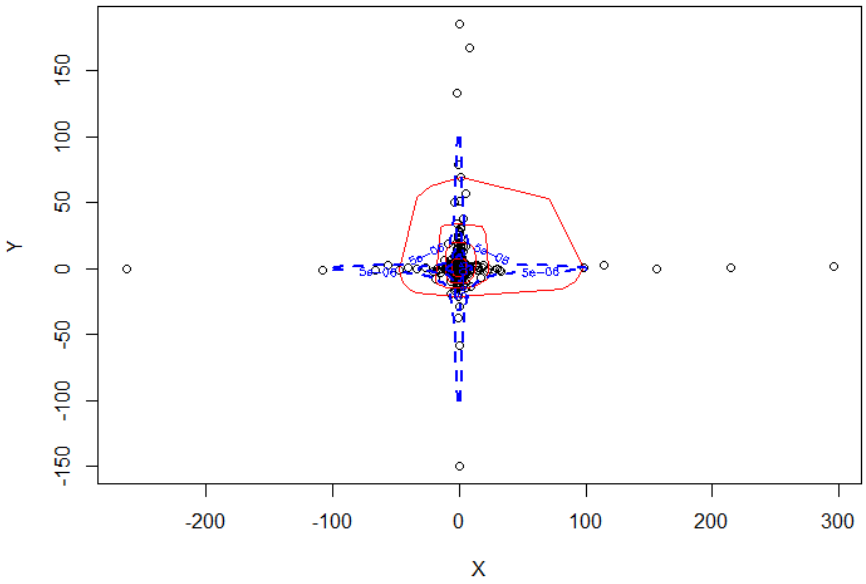

Figure 6: 1000 random samples from the bivariate Student's t-distribution with independent variables. Halfspace depth contours are drawn in red, isodensities in blue.

3.3 Higher dimension simulation

We want to make sure that depth functions still work in higher dimensions. Take for instance $d = 5$ and suppose that $X_i \in \mathbb{R}^5$ and are distributed $X_i \sim \mathcal{N}_5(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = \mathbf{1}$ and $\Sigma = I$ for all $1 \leq i \leq n$, where $n = 5000$ is the number of samples. The halfspace depth and density can be computed for every observation.

The depth range is $D_H \in [2 \cdot 10^{-4}; 0.436]$, which is nicely in line with Lemma 1. Moreover, if the mean of the distribution is computed, i.e. $\mathbf{1}$, then the depth is 0.4838. This is almost in line with Theorem 2. Practically, it is impossible to reach the maximum of 0.5, because a number of random samples are drawn, so it seems to be right that it is almost reached. The density range is $f \in [1.033 \cdot 10^{-9}; 0.010]$. Although we cannot compare these two ranges, we can however modify the depth range (the density range as well, but this is less natural) such that the two ranged coincide. This follows from the known fact that depth contours coincide with normal density contours. From Theorem 7, it follows that if $D_H(\mathbf{x}) = \alpha$, then for $d = 5$:

$$f(\mathbf{x}|\boldsymbol{\mu} = \mathbf{1}, \Sigma = I) = \frac{\exp(-\frac{1}{2}(\Phi^{-1}(1-\alpha))^2)}{(2\pi)^{5/2}}.$$

In figure 7 this relationship is plotted. All points follow the line $y = x$, so it follows that the contours would indeed coincide.

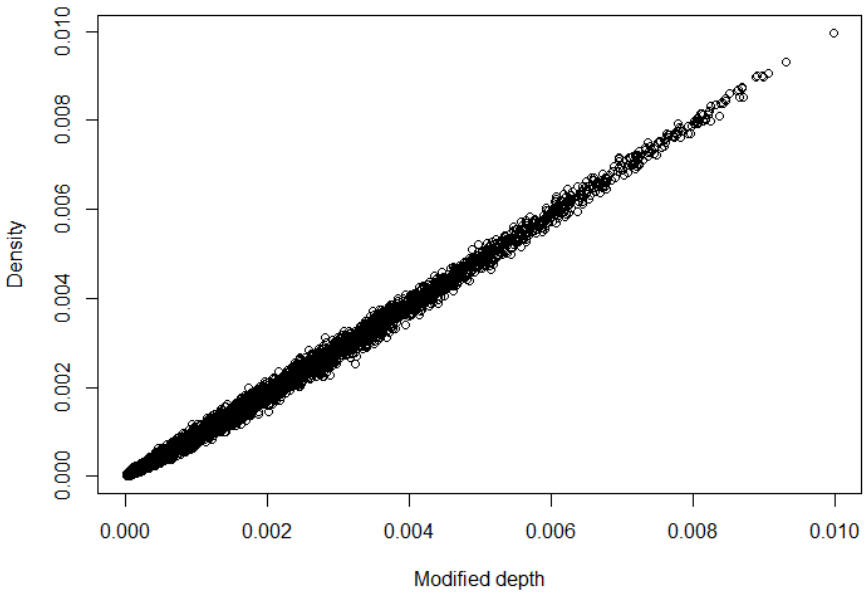

Figure 7: Density vs. depth of 5000 random normal fifthdimensional samples. The depth is modified such that it should coincide with the density.

4 Application in the weather

In this section the theory from Section 2 is applied to real-world data, specifically to data of the weather in the Netherlands. This data can be attained through the KNMI (Royal Netherlands Meteorological Institute), who keep track of all data regarding the weather. The KNMI has weather stations all across the country, with its headquarters in De Bilt. All data used in this section is about the weather station in De Bilt, see [15] for this data.

4.1 Temperature in July

The first data set to be researched concerns the temperature. The KNMI has been measuring the temperature since 1901 in De Bilt. The minimal, maximal and average temperature on every day since 1901 is available. The total amount of days from 1901 to 2018 sums up to 43099. Only the maximal temperature on a day will be researched in this report.

Working only with daily temperatures would mean that univariate data is researched. To apply the data to the theory in this report, it needs to be converted to multivariate data. One could see every year as an observation, which means the data would have a dimension of 365. There are two reasons why this is not a good idea. First of all, since we have a leap year every four years, this would mean that not every observation has the same dimension. Secondly, the dimension is too high to work with (this will be reasoned later on).

A second idea would be to work only with temperatures of a specific month. One of the motivations to write this report was to look into the July temperature of 2018, and find out how outlying it would be compared to the other July months. The dimension will now be 31, without observational exceptions, so every day of July is a variable.

The first thing to do after collecting the data is to make a quick analysation. Let $T(y, m, d)$ denote the maximum temperature on year y, month m and day d and let $T(y, m)$ denote the average maximum temperature of year y and month m, both in centigrades. In figure 8, the average maximum temperature of July is plotted against every year with a smooth curve. This curve indicates a rise in this kind of temperature.

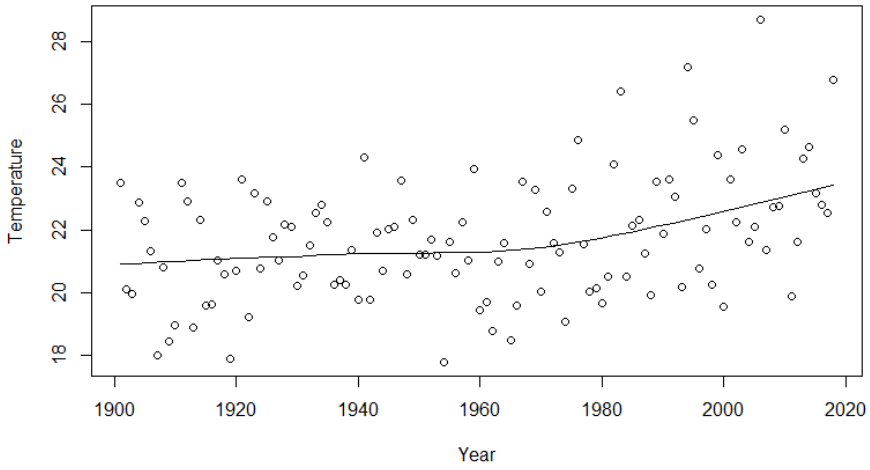

Figure 8: Average maximum temperature (in $°C$) of July plotted against every year, with a smooth curve, which denotes the trend.

The highest average maximum temperature was attained in 2006, with $T(2006, 7) = 28.7$. 2018 is the third highest, with $T(2018, 7) = 26.8$, and the lowest was attained in 1954, with $T(1954, 7) = 17.8$. All maximum temperatures in July range from $T(1903, 7, 7) = 11.1$ to $T(2006, 7, 19) = T(2018, 7, 26) = 35.7$.

Unfortunately. it is not possible in R to compute the simplicial depth for dimensions higher than 2, so only the halfspace depth will be used. However, the halfspace depth brings problems too. When computing the halfspace depth for every year with every day in July as a variable, all except four have the same value: $\frac{1}{118} \approx 0.0085$. Note that in this case we have used 118 observations of dimension 31. Obviously this is not a desired result: the dimension is still too high, and moreover the number of observations is too little for this dimension. Another idea is required.

The third and last idea is to investigate smaller parts of July. We will look into observations of dimension 7: exactly a week. There are 25 such weeks in July: 1 to 7, 2 to 8 etc. until 25 to 31. So actually we now have 25 data sets, and for every year the depth can be computed 25 times: all corresponding to some week in July. All these depths together range from 0.0085 to 0.27. The minimum is as expected, because it is actually 1/118. The maximum is also inside the range of 0 and 0.5, and 0.5 would almost never be reached because of the small number of observations.

There are four years which will be investigated. The first three are mentioned before: 1954 (lowest $T(y, 7)$), 2006 (highest $T(y, 7)$) and 2018 (third highest $T(y, 7)$). Another year of interest would be a very common year. One could pick the year for which the range in maximum temperature for all days is the smallest: it turns out to be 1974.

The depths range from 0.0085 to 0.11 for these four years. When visualizing the depths for these years, it is useful to use the outlyingness function, as defined in Subsection 2.5, since the depth is very dense and therefore the outlyingness will be more distinguishable. In figure 9 the outlyingness for four years is plotted against the 25 7-day intervals in July.

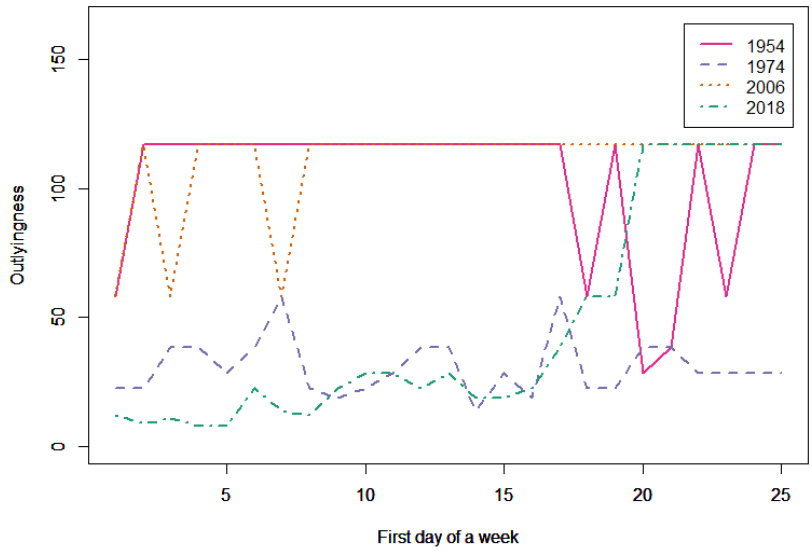

Figure 9: Maximum temperature (in $°C$) outlyingness plotted against 25 7-day intervals of July, for the years 1954, 1974, 2006 and 2018. A number on the x-axis corresponds to the week starting on that number.

One can see that the years 1954 and 2006 are more outlying than the other years. What does a high outlyingness mean in this case? It corresponds to a small depth, which means that the data is less deep or central in the data, and thus it is less common. But that does not mean that we can exclude such a week from being relatively cold or hot. Either one of the two is true. Indeed, in 1954 it was relatively cold, and in 2006 it was relatively hot, as can be established by looking to their mean and standard deviation, see table 1.

Furthermore, the years 1974 and 2018 seem to be a bit common, in the sense that they have lower outlyingness and higher depth. For 1974, this is true for all 7-day intervals. Although the mean of this year's maximum temperatures is not same as the mean of all year's maximum temperatures, it is close and moreover, its standard deviation is very low. This means that most maximum temperatures lie around the mean and therefore do not deviate much from the mean, so we would expect it to have a higher depth. For 2018, most of the 7-day interval depths have relatively high depth, except for the last part. If we look into the temperatures there, one can see that it has high outliers on the 26th and 27th of July, as they both are higher than 35.

One last note is that a lot of outlyingness numbers are 117. This corresponds to the lowest depth value of $\frac{1}{118}$.

Table 1: Mean and standard deviation of the maximum temperatures (in $^\circ C$) in July for four years.

Year	Mean	Standard deviation
1954	17.80323	1.905515
1974	19.09032	1.642428
2006	28.68065	3.107241
2018	26.77742	3.339831
All	21.71178	3.847131

4.2 Precipitation and wind speed

The second data set to be researched concerns both the precipitation and the wind speed. To be specific, for precipitation the highest amount of precipitation in an hour of the day is used, and for the wind speed the highest average wind speed of an hour of the day is used. With these variables, severe weather (e.g. heavy storm and rainfall) can be researched. The KNMI has been measuring the wind speed since 1904 and the precipitation since 1906 in De Bilt. As we will work with both data, only data from 1906 will be researched. From this year to 2018 corresponds with 41242 days.

At first, we will work with the bivariate dataset with 41242 observations. Obviously, this is a lot of data so there are limitations, which will be discussed later on. Again, there will be a quick analysation of the data. In figure 10, the highest hourly precipitation is plotted against the highest average wind speed in an hour for all days. In table 2, one can find the five dates with the lowest halfspace depth. These are also plotted in figure 10, indicated as red stars. The green star indicates the point with the highest depth.

Let $\beta(y, m, d) = (w(y, m, d), p(y, m, d))$ denote the observation of the highest average wind speed w in an hour and the highest amount of percipitation p in an hour on year y, month m and day d. Let $D(\beta)$ denote the halfspace depth of $\beta(w, p)$. A first thing to notice about figure 2 is that the wind speed data is discrete (at least more discrete than precipitation). Next, some ranges are given:

- $w \in [0.0; 26.8]$ (in m/s);

- $p \in [0.0; 44.1]$ (in mm);
- $D \in [2.425 \cdot 10^{-5}; 0.404]$.

All boundaries of the ranges of w and p are contained in table 2. The depth range is within the theoretical range (maximum of 0.5) as stated in Lemma 1. The minimal depth is in fact $\frac{1}{n}$, where $n = 41242$ is the number of observations. The maximal depth occurs 91 times in the point $\mu = (5.7; 0.3)$, hence μ is the halfspace depth median. There are observations worth noting. Firstly, $\beta(1962, 12, 5) = (0.0; 0.0)$, which is the only observation with $w = 0$. It means that on every hour of that day, the average wind speed was 0 m/s. Another notable observation is $\beta(1953, 6, 13) = (4.6; 44.1)$, which has the highest value of p. So on the 13th of June in 1953, there was an hour for which there was 44.1 mm precipitation. If one compares this with the fact that the average precipitation in De Bilt in June is 66 mm [16], it is quite shocking.

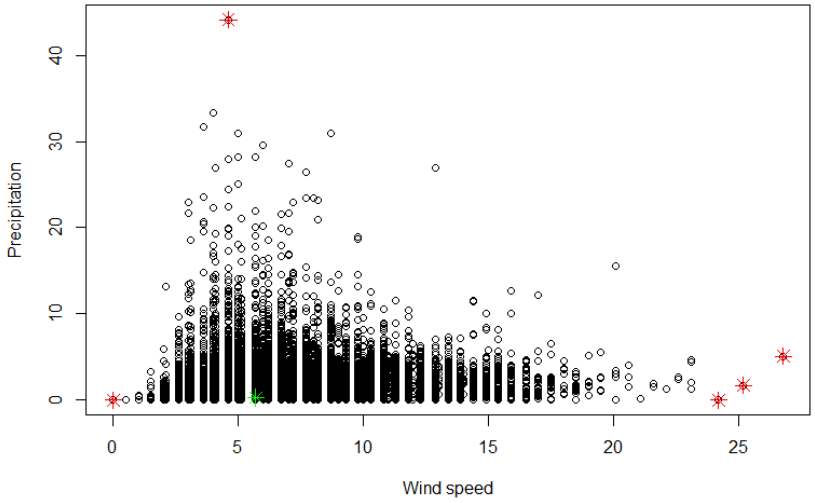

Figure 10: Highest amount of precipitation in an hour (in mm) vs. the highest average wind speed in an hour (in m/s) for all days since 1906. The smallest halfspace depths are indicated in red and the highest halfspace depth is indicated in green.

Table 2: Dates, wind speed, precipitation and depth for all observations with the highest halfspace depth. Data used since 1906.

Year	Month	Day	Wind speed (m/s)	Precipitation (mm)	Depth
1921	11	6	26.8	5.1	$2.425 \cdot 10^{-5}$
1930	1	13	24.2	0.0	$2.425 \cdot 10^{-5}$
1949	3	1	25.2	1.7	$2.425 \cdot 10^{-5}$
1953	6	13	4.6	44.1	$2.425 \cdot 10^{-5}$
1962	12	5	0.0	0.0	$2.425 \cdot 10^{-5}$

It is desirable to draw contours in this data set. However, because of the high number of observations, it takes statistical programming language R too long to compute these contours. Therefore, we need to work with less observations. To this end, let's take observations from 2012 onwards. There are now 2557 observations left.

In figure 11, the highest hourly precipitation is plotted against the highest average wind speed in an hour for all days since 2012. In table 3, one can find the five dates with the highest

halfspace depth.

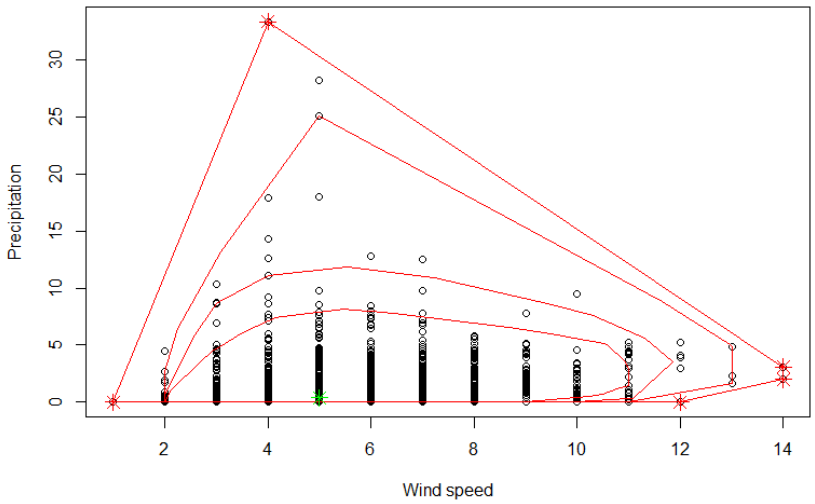

Figure 11: Highest amount of precipitation in an hour (in mm) vs. the highest average wind speed in an hour (in m/s) for all days since 2012. The smallest halfspace depths are indicated in red and the highest halfspace depths are indicated in green. The red lines are the halfspace depth contours.

Table 3: Dates, wind speed, precipitation and depth for all observations with the highest halfspace depth. Data used since 2012.

Year	Month	Day	Wind speed (m/s)	Precipitation (mm)	Depth
2013	10	28	14	3.1	$3.91 \cdot 10^{-4}$
2014	2	15	12	0.0	$3.91 \cdot 10^{-4}$
2016	6	23	4	33.3	$3.91 \cdot 10^{-4}$
2016	12	3	1	0.0	$3.91 \cdot 10^{-4}$
2018	1	18	14	2.0	$3.91 \cdot 10^{-4}$

Ranges are:

- $w \in [1; 14]$ (in m/s);
- $p \in [0.0; 33.3]$ (in mm);
- $D \in [0.0003910833; 0.3793508]$.

All boundaries of the ranges of w and p are again contained in table 3. The depth range is within the theoretical range (maximum of 0.5) as stated in Lemma 1. The minimal depth is in fact $\frac{1}{n}$, where $n = 2557$ is the number of observations. The maximal depth occurs 12 times in the point $\mu = (5; 0.4)$, hence μ is the halfspace depth median.

Because the number of observations is now more passible, halfspace depth contour lines can be computed. These are also implemented in figure 11, with levels $\alpha \in \frac{1}{n}\{1, 3, 9, 20\}$. The contour lines provide useful information about the distribution of the data. With just a twodimensional plot, it is not possible to see how much data there is near low precipitations (i.e. near zero precipitation). Since the contours are so close to each other at the bottom, it must be

the case that the data is extremely dense near 0 precipitation. Furthermore, the contours are nested within each other, which is in line with Theorem 4. Note that the wind speed data being discrete has influence on the contour shapes. The left and right edges of the contours are near the discrete observations.

5 Conclusion

The main goal of this report was to find a way to measure data depth in multivariate data. Depth functions provide a solution to this problem. We desire from these depth function that they order data from central observations to outliers. The halfspace depth of a point is the smallest ratio of points in some halfspace which contains that point. For central points, there are a lot of points around, so every halfspace contains lots of points, and therefore it has high depth. Outliers are situated far away from other data, so there probably exists a halfspace for which there are no other observations, and therefore it has low depth. The simplicial depth of a point is the ratio of simplices (with observations as vertices) which contain that point. Central points are surrounded by all other points, so they are contained in a lot of simplices, and therefore has high depth. Outliers are probably only contained in simplices for which the outlier is a vertex itself, so it has low depth. In conlusion, both depth functions provide a good notion of ordering multivariate data.

But which of these is better? Theoretically speaking, none of them is better than the other. To illustrate, we have looked into the differences between the simplicial and halfspace depth for the bivariate normal distribution. It turns out that the ordering is almost the same, with some rotation between a couple of observations. Their order did not differ much from each other. Practically speaking, the halfspace depth is better. First of all, this depth function is more common, so it has less limitations in programming language R than the simplicial depth function. Moreover, we saw that for the bivariate normal distribution, the contours are less shaped as circles/ellipses than the halfspace depth.

Finally, some real-world data was researched. For the maximal temperature, we have looked into four specific years: 1954 (lowest T), 1974 (smallest T range), 2006 (highest T) and 2018 (third highest T). Both 1954 and 2006 had high outlyingness, and 1974 had a relatively low outlyingness, so far as expected. However, 2018 had a high depth for the first 15 7-day intervals. This is not really expected as 2018 had the third highest average maximum temperature. This is a consequence of the extreme high temperatures on the last 10 7-day intervals. We have also looked into the combination of heavy wind and precipitation. The depth is computed and some outliers originated. With the depth contours, we can qualify how outlying this data is, and therefore it gives an idea of how to define severe weather. Moreover, the frequency of severe weather can be estimated. This is something which can be investigated more in another research.

6 Discussion

During the process of this research, there were some difficulties. Firstly, the depth functions could not be expressed in terms of distribution functions for higher dimensions. If these were known, then computing the depth would be a lot easier. For the univariate case however, they were found and in fact have been proven in this report. In the computation of the depth of bivariate normal observations, it followed that the highest simplicial depth exceeded the maximum of 0.25. I propose two reasons for this: there are not enough observations and the computation of the simplicial depth is not 100% waterproof. Obviously, if there are not enough observations, then computations are less exact. However, if more observations were used, then the simplicial depth contained negative values and values higher than 30 for instance. Therefore, I think that the simplicial depth function in R is not completely correct. Another notable phenomenon was found in the case of dependent bivariate normal variables. The halfspace depth contours, from outside to inside, first seem to coincide more and more with the isodensities, but from the inner contour on they just coincided less. I have not yet figured out why this occurs, so this is definitely something for future research. Also, I wanted to create contours for threedimensional observations. These would be spheres. However, I was not able to do so, as I have tried many functions in R, but unfortunately none of them worked. Furthermore, for the temperature data, the idea was to compute the depth for every year, where every year was an observation of all days in July. That means that the dimension would be 31, with 118 observations. The dimension unfortunately was too high to make accurate calculations. The conclusion that 2018 was not very outlying was alsog not very satisfying, because it was the year with the third highest average maximum temperature of July. The reason is that there are not enough observations. With more observations, the last ten intervals could have more potential for high outlyingness than other years. Finally, the depth orders data and specifically detects outliers. In our real-world data, we expect outliers to be extreme weather situations. However, when computing the depth there are mild weather situations with very low depth, in which we are not interested. Think of the lowest July temperature, or the observation of zero average wind speed in an hour.

References

[1] Yorick de Wijs. *Juli 2018*. July 2018. URL: https://www.knmi.nl/nederland-nu/klimatologie/maand-en-seizoensoverzichten/2018/juli.

[2] Sigrid Landman. *De droge zomer van 2018; wat zijn de gevolgen*. Dec. 2018. URL: https://wetenschap.infonu.nl/weer/189941-de-droge-zomer-van-2018-wat-zijn-de-gevolgen.html.

[3] John W. Tukey. "Mathematics and picturing data". In: *Proceedings of the International Congress on Mathematics*. Vol. 2. 1975, pp. 523–531.

[4] Regina Y. Liu. "On a notion of data depth based on random simplices". In: *The Annals of Statistics* 18.1 (1990), pp. 405–414.

[5] Robert Serfling. "Depth Functions in Nonparametric Multivariate Inference". In: *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 72 (2006), pp. 1–16.

[6] Yijun Zuo and Robert Serfling. "Structural properties and convergence results for contours of sample statistical depth functions". In: *The Annals of Statistics* 28.2 (2000), pp. 483–499.

[7] Peter J. Rousseeuw and Ida Ruts. "The depth function of a population distribution". In: *Metrika* 49 (1999), pp. 213–244.

[8] Michael C. Hughes. *Why probability contours for the multivariate Gaussian are elliptical*. 2013. URL: https://www.michaelchughes.com/blog/2013/01/why-contours-for-multivariate-gaussian-are-elliptical/.

[9] Karl Mosler. *Depth statistics*. Cologne, Germany: Universität zu Köln, 2012.

[10] Chuong B. Do. *The multivariate Gaussian Distribution*. Stanford, CA: Stanford University, 2008.

[11] Yu Tsumura. *Positive definite Real Symmetric Matrix and its Eigenvalues*. URL: https://yutsumura.com/positive-definite-real-symmetric-matrix-and-its-eigenvalues/.

[12] Michael Roth. *On the Multivariate t Distribution*. Linköping, Sweden: Linköpings universitet, 2013.

[13] Christopher M. Bishop. "Pattern Recognition and Machine Learning". In: *Springer* (2006).

[14] Daniel T. Cassidy. "A Multivariate Student's t-Distribution". In: *Open Journal of Statistics* 6 (2016), pp. 443–450.

[15] KNMI. *Daggegevens van het weer in Nederland*. June 2019. URL: http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi.

[16] Carine Homan. *Juni 2018*. July 2018. URL: https://www.knmi.nl/nederland-nu/klimatologie/maand-en-seizoensoverzichten/2018/juni.

A Proofs

Beneath, the proofs of lemmas and theorems can be found.

Theorem 1. *If F is a distribution in $\mathbb{R}^1$, then*

$$D_H(z) = \min\{F(z), 1 - F(z)\} \in \left[0, \frac{1}{2}\right].$$

Moreover, the depth is maximized in the median μ with maximum value $\frac{1}{2}$.

Proof. For computation of the halfspace depth, it may be assumed that the boundary of a closed halfspace H which contains z and minimizes $P(H)$ passes through that point z. Thus, for the univariate case, only two halfspaces are left: $\{x \in \mathbb{R}^1 |\ x \leq z\}$ and $\{x \in \mathbb{R}^1 |\ x \geq z\}$. The probabilities of these halfspaces are respectively $F(z)$ and $1 - F(z)$, and therefore we have that

$$D_H(z) = \min\{F(z), 1 - F(z)\}.$$

Now, since $F(z) \in [0, 1]$, it immediately follows that $D_H(z) \in \left[0, \frac{1}{2}\right]$. Moreover, since $F(\mu) = \frac{1}{2}$ per definition, it follows that $\max_z D_H(z) = D_H(\mu) = \frac{1}{2}$. □

Lemma 1. *For any probability P with a density we have $D_H(z) \leq \frac{1}{2}$ for any $z \in \mathbb{R}^d$.*

Proof. Let H_z be some closed halfspace with its boundary passing through z and let $H_z^- := H_z^c \cup \partial H_z$. Then H_z^- is a closed halfspace with its boundary passing through z as well, and $H_z \cap H_z^- = \partial H_z$, so this is the hyperplane passing through z. Then we have

$$P(H_z) + P(H_z^-) = P(H_z \cup H_z^-) + P(H_z \cap H_z^-) = 1,$$

since the probability of a hyperplane is zero (because P has a density). From this, it follows that $\min\{P(H_z), P(H_z^-)\} \leq \frac{1}{2}$. At any z we now have

$$D_H(z) = \inf\ P(H_z) = \inf \min\{P(H_z), P(H_z^-)\} \leq \frac{1}{2}$$

□

Theorem 2. *If P has a density and is angularly symmetric about μ, then*

$$\max_z D_H(z) = D_H(\mu) = \frac{1}{2}.$$

Proof. By Lemma 1 it is sufficient to show that $D_H(\mu) = \frac{1}{2}$. Because P has a density, the hyperplane ∂H_μ has zero probability. Therefore, $P(H_\mu) = P(\text{int}(H_\mu))$. By angular symmetry, it follows that $P(\text{int}(H_\mu)) = P(\text{int}(H_\mu^-))$ and hence $P(H_\mu) = P(H_\mu^-)$. Both equal $\frac{1}{2}$ by the proof of lemma 1. This is true for every halfspace passing through μ, hence $D_H(\mu) = \frac{1}{2}$. □

Theorem 3. *Let $X_1, X_2 \in \mathbb{R}^1$ be i.i.d. with univariate continuous cdf F and let $\overline{X_1 X_2}$ be the line segment connecting X_1 and X_2. The simplicial depth will then be*

$$D_S(x|F) = P_F(x \in \overline{X_1 X_2}) = 2F(x)(1 - F(x)).$$

Moreover, the depth is maximized in the median with maximum value $\frac{1}{2}$.

Proof. Define the events $A := X_1 \leq x \leq X_2$ and $B := X_2 \leq x \leq X_1$. We have that $\{x \in \overline{X_1 X_2}\} \iff A \cup B$, hence

$$P(x \in \overline{X_1 X_2}) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - 0 = 2P(A)$$

The latter equation follows from X_1 and X_2 being identically distributed. For the probabilty of event A we have

$$P(A) = P(X_1 \leq x \leq X_2) = P(X_1 \leq x,\ x \leq X_2) = P(X_1 \leq x)P(x \leq X_2),$$

since X_1 and X_2 are independent. Finally, because $X_1, X_2 \sim F$, it follows that

$$D(x) = 2P(X_1 \leq x)(1 - P(X_2 \leq x)) = 2F(x)(1 - F(x)).$$

Now that the simplicial depth is expressed in terms of F, the deepest point in the data set can explicitly be found. The deepest point, or the median μ, is the observation for which D is maximized, thus

$$\mu := \arg\max_{x \in \mathbf{R}^1} D(x) = \arg\max_{x \in \mathbf{R}^1} 2F(x)(1 - F(x))$$

One way to maximize this is through differentiating:

$$D'(x) = 2F'(x)(1 - F(x)) + 2F(x) \cdot -F'(x) = 2f(x)(1 - 2F(x))$$

$$D'(x) = 0 \iff 1 - 2F(x) = 0 \iff F(x) = \frac{1}{2}.$$

Note that $f(x) = F'(x)$ is the density function of $F(x)$. Per definition, $F(\mu) = \frac{1}{2}$ for median μ, so $D(x)$ is indeed maximized by the median. In conclusion, $\max_{x \in \mathbb{R}^1} D(x) = D(\mu) = \frac{1}{2}$. □

B Data

Table 4: Random samples from $\mathcal{N}_2(\boldsymbol{\mu}_1, \Sigma_1)$ with their halfspace and simplicial depths

x_i	y_i	$D_H(z_i)$	$D_S(z_i)$
-0.35615523	4.00089943	0.001	0.003000000
1.71681764	3.87372861	0.001	0.003000000
-1.43443828	3.68583033	0.001	0.003000000
3.83327294	3.03635224	0.001	0.003000000
4.23115697	1.98202156	0.001	0.003000000
-2.00492196	0.63533572	0.001	0.003000000
4.17306860	-0.72755655	0.001	0.003000000
-1.23203868	-1.76603419	0.001	0.003000000
2.13909391	-1.93266328	0.001	0.003000000
0.99863469	-1.93578445	0.001	0.003000000
⋮	⋮	⋮	⋮
1.0016427	1.0915449	0.439	0.2499906
0.8878813	1.0727613	0.443	0.2484072
0.8985231	1.0031238	0.449	0.2509111
0.9593943	1.0691585	0.455	0.2509392
0.9653366	0.9019999	0.456	0.2509568
0.9402226	0.8940368	0.456	0.2503058
1.0053278	1.0005656	0.470	0.2524526
0.9709936	0.9425479	0.471	0.2522974
0.9671003	0.9500382	0.473	0.2524496
0.9899894	0.9767099	0.481	0.2526816

Table 5: Random samples from $\mathcal{N}_2(\boldsymbol{\mu}_2, \Sigma_2)$ with their halfspace and simplicial depths

x_i	y_i	$D_H(z_i)$	$D_S(z_i)$
2.10877815	4.54321643	0.001	0.003000000
2.92077753	4.27693276	0.001	0.003000000
3.84713079	3.13031316	0.001	0.003000000
-0.81826951	2.18665245	0.001	0.003000000
4.18016925	1.34689630	0.001	0.003000000
3.46603410	0.23487713	0.001	0.003000000
-2.51147522	-0.27943654	0.001	0.003000000
-1.54314657	-1.54178126	0.001	0.003000000
1.09042644	-2.08264216	0.001	0.003000000
-0.97021394	-2.10930786	0.001	0.003000000
⋮	⋮	⋮	⋮
1.0801015	1.0784589	0.439	0.2499906
1.0069538	1.1190725	0.443	0.2484072
0.9519668	1.0534437	0.449	0.2509111
1.0395902	1.0801959	0.455	0.2509392
0.8783445	0.9381219	0.456	0.2503058
0.8977977	0.9324611	0.456	0.2509568
1.0031537	0.9978259	0.470	0.2524526
0.9357418	0.9647482	0.471	0.2522974
0.9402820	0.9731817	0.473	0.2524496
0.9748249	0.9848355	0.481	0.2526816

C Code in R

Simulation with multivariate distributions

```
library(abind);library(depth);library(dplyr);library(MASS);library(Matrix);
library(mvtnorm);library(plot3D);library(plotly);library(RColorBrewer);
library(reshape2);library(tidyr);library(laGP); library(ContourFunctions);
library(fMultivar);
# Bivariate normal distribution 1, independent variables
set.seed(127)
n <- 1000
mu <- c(1,1)
Sigma <- diag(2)
x <- mvrnorm(n, mu, Sigma)
dpt_tukey <- numeric(n)
for (i in 1:n){dpt_tukey[i] <- depth(x[i,],x,method="Tukey")}
dpt_liu <- numeric(n)
for (i in 1:n){dpt_liu[i] <- depth(x[i,],x,method="Liu")}
md <- data.frame(v1=x[,1],v2=x[,2],tukey_depth=dpt_tukey,liu_depth=dpt_liu)
outliers <- md %>% filter(tukey_depth==min(md$tukey_depth))
non_outliers <- filter(md, row_number(desc(tukey_depth)) <= 10)

#1e plot
plot(md[,1:2],xlab="X",ylab="Y")
points(outliers$v1,outliers$v2,pch=8,cex=1.5,col="red")
points(non_outliers$v1,non_outliers$v2,pch=8,cex=1.5,col="green")
abline(h=mean(md[,2]),v=mean(md[,1]),lwd=1,col="blue")

#2e plot
isodepth(md[,1:2],dpth=c(1,5,20,50,100),colcontours="Red")
#isodensities
N <- 100
x.points <- seq(-4,6,length.out=N)
y.points <- x.points
z <- matrix(0,nrow=N,ncol=N)
for (i in 1:N) {for (j in 1:N) {z[i,j] <-
    dmvnorm(c(x.points[i],y.points[j]),mean=mu,sigma=Sigma)}}
lvls = 1/n * c(1,5,20,50,100)
adj_lvls <- exp(-(qnorm(1-lvls))^2/2)/(2*pi)
contour(x.points,y.points,z,levels=adj_lvls,lty=2, lwd=2,add=TRUE,col="blue")
Liu_dpt <- perspdepth(md[,1:2],method="Liu",output=TRUE)
contour(Liu_dpt[["x"]],Liu_dpt[["y"]],Liu_dpt[["z"]],
    levels=lvls,lwd=2,add=TRUE,col="green3")

# Bivariate normal distribution, dependent variables
set.seed(127)
n <- 1000
mu <- c(1,1)
rho <- 0.5
```

```
Sigma <- matrix(c(1,rho,rho,1),2,2)
x <- mvrnorm(n, mu, Sigma)
dpt_tukey <- numeric(n)
for (i in 1:n){dpt_tukey[i] <- depth(x[i,],x,method="Tukey")}
dpt_liu <- numeric(n)
for (i in 1:n){dpt_liu[i] <- depth(x[i,],x,method="Liu")}
md <- data.frame(v1=x[,1],v2=x[,2],tukey_depth=dpt_tukey,liu_depth=dpt_liu)
outliers <- md %>% filter(tukey_depth==min(md$tukey_depth))
non_outliers <- filter(md, row_number(desc(tukey_depth)) <= 10)

#1e plot
plot(md[,1:2],xlab="X",ylab="Y")
points(outliers$v1,outliers$v2,pch=8,cex=1.5,col="red")
points(non_outliers$v1,non_outliers$v2,pch=8,cex=1.5,col="green")
abline(h=mean(md[,2]),v=mean(md[,1]),lwd=1,col="blue")

#2e plot
isodepth(md[,1:2],dpth=c(1,5,20,50,100),colcontours="Red")
#isodensities
N <- 100
x.points <- seq(-4,6,length.out=N)
y.points <- x.points
z <- matrix(0,nrow=N,ncol=N)
for (i in 1:N) {for (j in 1:N) {z[i,j] <-
    dmvnorm(c(x.points[i],y.points[j]),mean=mu,sigma=Sigma)}}
lvls = 1/n * c(1,5,20,50,100)
adj_lvls <- exp(-(qnorm(1-lvls))^2/2)/(2*pi*sqrt(det(Sigma)))
contour(x.points,y.points,z,levels=lvls,lty=2, lwd=2,add=TRUE,col="blue")
Liu_dpt <- perspdepth(md[,1:2],method="Liu",output=TRUE)
contour(Liu_dpt[["x"]],Liu_dpt[["y"]],Liu_dpt[["z"]],
    levels=lvls,lwd=2,add=TRUE,col="green3")

# dimension 5

set.seed(127)
n <- 5000
mu <- rep(1,5)
Sigma <- diag(5)
x <- mvrnorm(n, mu, Sigma)
head(x)
dpth5 <- numeric(n)
dens5 <- numeric(n)
for (i in 1:n){dpth5[i] <- depth(x[i,1:5],x[,1:5],approx=TRUE)}
for (i in 1:n){dens5[i] <- dmvnorm(x[i,1:5],mean = mu,sigma = Sigma)}
matr5 <- cbind(x,dpth5,dens5)
depth(numeric(5)+1,x[,1:5],approx = TRUE)
plot(dpth5,dens5,xlab="Depth",ylab="Density")
dpth5_mod <- exp(-(qnorm(1-dpth5))^2/2)/(2*pi)^(5/2)
plot(dpth5_mod,dens5,xlab="Modified depth",ylab="Density")
```

```
#student t distribution

set.seed(127)
n <- 1000
nu <- 1
x <- rt(n, df = nu)
set.seed(141)
y <- rt(n, df = nu)
isodepth(cbind(x,y),dpth=c(5,10,20,50),col="Red")

N <- 100
x.points <- seq(-100,100,length.out=N)
y.points <- x.points
z <- matrix(0,nrow=N,ncol=N)
for (i in 1:N) {for (j in 1:N) {z[i,j] <-
    1/(pi^2*(1+x.points[i]^2)*(1+y.points[j]^2))}}
lvls = 1/200000*c(1,10,100)
contour(x.points,y.points,z,levels=lvls,lty=2, lwd=2,add=TRUE,col="blue")
```

Application on temperature

```
    setwd("D:/Documenten/TU Delft schoolwerken/Bacherlorproject") #workspace
x <- read.table("KNMI_all_temp.txt",sep = ",")
n <- nrow(x)
yr <- 1:n; mn <- 1:n; dy <- 1:n; temp <- 1:n;
for (i in 1:n){
  yr[i] <- strtoi(substring(x[i,2],1,4),base=10)
  mn[i] <- strtoi(substring(x[i,2],5,6),base=10)
  dy[i] <- strtoi(substring(x[i,2],7,8),base=10)
  temp[i] <- x[i,3]/10
  }
mydata <- data.frame(year = yr, month = mn, day = dy, temperature = temp)

#maandgemiddelde van input(maand) voor elk jaar, trendlijn
maand = "Jul" #first three letters of the month, with first letter as capital
yrs <- 1901:2018; month_avg <- 1:118
for (i in yrs){
  workdata <- subset(mydata, year==i & month ==match(maand,month.abb),
    select=temperature)
  month_avg[i-1900] <- mean(workdata[,])}
#data analyseren
scatter.smooth(yrs,month_avg,xlab="Year",ylab="Temperature")
July_data <- subset(mydata, month ==match(maand,month.abb))
View(July_data)
View(month_avg)

#depth berekenen met 118 observaties in dimensie 31: werkt niet
```

```
maand = "Jul"
data_wide <- spread(mydata[ which(mydata$month ==match(maand,month.abb)), ]
    , day, temperature)
depth_matrix <- as.matrix(data_wide[1:118,3:33]) #118 observaties van dimensie 31
depth_obs <- matrix(nrow=118,ncol=1)
for (j in 1:118){ depth_obs[j] <-
    depth(depth_matrix[j,],depth_matrix,approx = TRUE) }

#depth berekenen met 118 observaties in dimensie 7
maand = "Jul"
data_wide <- spread(mydata[ which(mydata$month ==match(maand,month.abb)), ]
    , day, temperature)
weekly_depth <- matrix(nrow=118,ncol=25)
for (i in 1:25){
  k <- i+2
  l <- i+8
  depth_matrix <- as.matrix(data_wide[1:118,k:l])
  for (j in 1:118){
    depth_point <- as.numeric(depth_matrix[j,])
    weekly_depth[j,i] <- depth(depth_point,depth_matrix,approx=TRUE)
  }
}
min(weekly_depth)
max(weekly_depth)

#mogelijkheid: plot voor elk interval (25 totaal) de diepte en verbind de punten.
#1: bepaal het jaar waarvoor geldt dat de range in max temperatuur
# in heel juli het laagst is (1974)
u<-numeric(118);
for (i in 1:118){u[i]<-max(data_wide[i,3:33])-min(data_wide[i,3:33])};
which.min(u)
weekly_outl <- 1/weekly_depth - 1
years <- c(1954,1974,2006,2018) #niet meer dan 6
k <- length(years)
colors <- rev(brewer.pal(k,"Dark2"))
M <- 1.4*max(weekly_outl)
for (i in 1:k){
  if (i == 1){
    yr <- years[i]-1900
    plot(xlab="First day of a week",ylab="Outlyingness",1:25,weekly_outl[yr,]
        ,ylim=c(0,M),col=colors[i],type = "l",lty=i,lwd=2)
    }
  else{
    yr <- years[i]-1900
    par(new=TRUE)
    plot(xlab="First day of a week",ylab="Outlyingness",1:25,weekly_outl[yr,]
        ,ylim=c(0,M),col=colors[i],type = "l",lty=i,lwd=2)
    }
}
```

```
legend(x=21.5,y=M,legend=years,col=colors,lty=1:k,lwd=2)

#lijkt dit te kloppen?
for (k in years){
  print(k)
  print(mean(July_data[July_data$year==k,4]))
  print(sd(July_data[July_data$year==k,4]))
}
mean(July_data[,4])
sd(July_data[,4])
```

Application on precipitation and wind speed

```
  #data formatteren
setwd("D:/Documenten/TU Delft schoolwerken/Bacherlorproject") #workspace
x <- read.table("KNMI_all_2.txt",sep = ",")
m <- nrow(x)
yr <- 1:m; mn <- 1:m; dy <- 1:m; wd <- 1:m; rn <- 1:m; dp <- 1:m
for (i in 1:m){
  yr[i] <- strtoi(substring(x[i,2],1,4),base=10)
  mn[i] <- strtoi(substring(x[i,2],5,6),base=10)
  dy[i] <- strtoi(substring(x[i,2],7,8),base=10)
  wd[i] <- x[i,3]/10 #unit: m/s
  rn[i] <- x[i,4]/10 #unit: mm
}
mydata <- na.omit(data.frame(year = yr, month = mn, day = dy, wind = wd
    , rain = rn, dpth = dp))
#wind vanaf 1904, regen vanaf 1906
n <- nrow(mydata)
for (i in 1:n){
  if (mydata[i,5]<=0){
    mydata[i,5]<-0 #er zijn dagen waarop de neerslag negatief is gezet ipv op 0
  }
}
#depth voor alle observaties:
n <- nrow(mydata)
for (i in 1:n){mydata[i,6] <- depth(mydata[i,c(4,5)],mydata[,c(4,5)])}
outliers1 <- mydata %>% filter(dpth==min(mydata$dpth))
non_outliers1 <- mydata %>% filter(dpth==max(mydata$dpth))
plot(mydata[,4:5],xlab="Wind speed",ylab="Precipitation")
points(outliers1$wind,outliers1$rain,pch=8,cex=1.5,col="red")
points(non_outliers1$wind,non_outliers1$rain,pch=8,cex=1.5,col="green")
#cf_data(mydata[,4],mydata[,5],mydata[,6])

#observaties vanaf 2012 (2557 obs)
md2000 <- mydata[mydata$year>2011,]
n <- nrow(md2000)
for (i in 1:n){md2000[i,6] <- depth(md2000[i,c(4,5)],md2000[,c(4,5)])}
```

```
outliers <- md2000 %>% filter(dpth==min(md2000$dpth))
non_outliers <- md2000 %>% filter(dpth==max(md2000$dpth))
isodepth(md2000[,4:5],dpth = c(1,3,9,20),mustdith = TRUE,colcontours="Red"
    ,xlab="Wind speed",ylab="Precipitation")
points(outliers$wind,outliers$rain,pch=8,cex=1.5,col="red")
points(non_outliers$wind,non_outliers$rain,pch=8,cex=1.5,col="green")
#perspdepth(md2000[,4:5])
```