FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Juraj Bodík

Geometric approach to the estimation of scatter

Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis:	Mgr. Stanislav Nagy, Ph.D.
Study programme:	Mathematics
Study branch:	General Mathematics

Prague 2019

I declare that I carried out this bachelor 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 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date .... signature of the author

I would like to thank my supervisor Mgr. Stanislav Nagy, Ph.D. for his comments and suggestions, for his great help during the whole semester, and specially for his patience with my bad grammar.
I would also like to thank my parents for their support in my whole life.

Title: Geometric approach to the estimation of scatter

Author: Juraj Bodík

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Stanislav Nagy, Ph.D.

Abstract: In this thesis we describe improved methods of estimating mean and scatter from multivariate data. As we know, the sample mean and the sample variance matrix are non-robust estimators, which means that even a small amount of measurement errors can seriously affect the resulting estimate. We can deal with that problem using MCD estimator (minimum covariance determinant), that finds a sample variance matrix only from a selection of data, specifically those with the smallest determinant of this matrix. This estimator can be also very helpful in outlier detection, which is used in many applications. Moreover, we will introduce the MVE estimator (minimum volume ellipsoid). We will discuss some of the properties and compare these two estimators.

Keywords: MCD estimator, MVE estimator, Mahalanobis distance, scatter matrix, robustness, breakdown point

Název práce: Geometrický přístup k odhadování rozptýlenosti

Autor: Juraj Bodík

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Stanislav Nagy, Ph.D.

Abstrakt: V tejto práci popisujeme vylepšené metódy na odhadovanie polohy a rozptýlenosti viacrozmerných dát. Výberový priemer a výberová rozptylová matica sú nerobustné metódy, čo znamená že aj jedno zlé pozorovanie môže tento odhad znehodnotiť. Tento problém rieši MCD odhad (minimum covariance determinant), ktorý spočíta strednú hodnotu a variačnú maticu iba z vhodnej selekcie dát, konkrétne z pozorovaní ktorých variačná matica má najmenší determinant. Vhodná aplikácia je v hľadaní odľahlých pozorovaní. Na záver ukážeme ďalší postup, a to MVE odhad (minimum volume ellipsoid). Budeme diskutovať ich vlastnosti a porovnáme tieto dva odhady.

Klíčová slova: MCD odhad, MVE odhad, Mahalanobisova vzdialenosť, matica rozptýlenosti, robustnosť, bod zlomu

Contents

Introduction

The classical mean vector and covariance matrix are the cornerstones in many multivariate statistical methods. Sometimes, we need to consider their robust alternatives, because they are extremely sensitive to outliers. Robustness is a very desired property of an estimator, representing the ability of an estimator to handle a variety of distributions, including outliers. Robust estimators of location and scatter are widely used in practice, e.g. in linear regression or principal component analysis [1]. In this thesis we will show one specific usage, and that is outlier detection.

Let $X_1, \ldots, X_n$ be a random sample from a bivariate normal distribution with the expected value μ and the variance matrix Σ. The question is: how does the region containing 99 % of the mass of this distribution look like? In other words, we want to find a set containing a new independent observation generated from this model with probability 99 %. One possible approach is to take a suitable level set of the density function, whose probability is 99 %. Every level set of the density function of a normal distribution forms an ellipse, therefore an appropriate ellipse will be our confidence region (also called a 99% tolerance ellipse). Moreover, this ellipse obviously depends only on μ and Σ, therefore if we know μ and Σ, we can draw it. But how can we proceed if we do not know μ and Σ? How can we find this confidence region only from data? This and more will be discussed in Chapter 1. The intuitive approach is to estimate μ and Σ with sample mean and sample variance matrix to compute the confidence region with them. It is a good method and we obtain reasonable results.

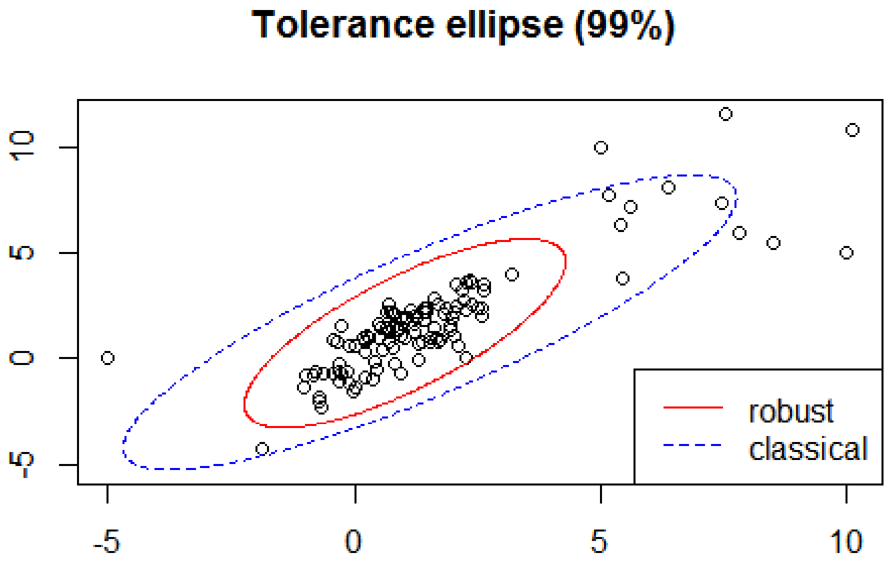


Figure 1: Importance of using robust estimators. The figure shows 99% tolerance ellipse for bivariate data based on robust and non-robust estimations of the expected value μ and the variance matrix Σ.

Now, this problem gets much more difficult if our data contains outliers. An outlying observation, or an outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. It is hard to define formally because an outlier can occur by chance in any distribution, but it often indicates a measurement error. With the outliers, estimates of μ and Σ using the sample mean and the sample variance are heavily damaged. Therefore, we need to find their robust alternative. In Figure 1 we can see the difference between computing the tolerance ellipse using robust and classical methods.

The core of the thesis is the MCD estimator (Minimum Covariance Determinant). It is an estimator of location and scatter, that is very robust. For a whole class of distributions (including normal), MCD estimator can estimate μ and Σ consistently regardless of outlying observations. The main idea is to use only a suitable subset of our data (not outliers), and compute μ and Σ only from this selection. The definition of the estimator is in Chapter 2. Of course, everything comes with a price — if we want a robust estimator, we will pay with other worse properties. In Chapter 3 we will find these properties and discuss possible improvements.

In Chapter 4, we will present other highly robust estimator called the MVE estimator (Minimum Volume Ellipsoid). This estimator uses mainly the geometric interpretation of a tolerance ellipse (or an ellipsoid in more dimensions). The main idea is the following. Instead of estimating μ and Σ to compute the tolerance ellipsoid, we will first find an appropriate tolerance ellipsoid to compute μ and Σ. We can try to fit the smallest possible ellipsoid containing some fraction of the data (that are not likely to be outliers) and inflate it, such that the ellipsoid will represent an estimate of the 99% tolerance ellipse. Then we can easily compute μ and Σ from its location and shape.

We will show that these estimators are the most robust estimators — there does not exist a “reasonable” estimator that can handle more outliers. Moreover, we will use these estimators in one application in Section 2.1.

0.0.1 Symbols, notations and definitions

In the entire thesis we assume the following.

Let $n, p \in \mathbb{N}$. Let $\mu \in \mathbb{R}^p$ and $\Sigma \in PDS(p)$ be unknown parameters ($PDS(p)$ denotes the class of positive definite symmetric matrices with dimension p). We consider $X_1, \ldots, X_n$ random sample from $N_p(\mu, \Sigma)$. Suppose we have data from this distribution, the i-th observation is denoted by $x_i = (x_{i,1}, ..., x_{i,p})^t$, and stored in a matrix $\mathbb{M} = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$.

List of all symbols and notations can be found in Attachment A. There, also some basic theorems from linear algebra are given that will be used in the thesis.

The most important notations are these: Matrices are denoted by bold capital letters such as $\mathbb{A}, \mathbb{B}, \mathbb{C}, \ldots$. Random vectors and variables have the same notation $X, Y, \ldots$. By default we consider vectors as column vectors, $x = (x_1, \ldots, x_n)^t$. The notation $X \sim N_p(\mu, \Sigma)$ stands for X is the p-dimensional random vector with normal distribution (with parameters $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$).

Usually, we consider only normal vectors (possibly contaminated by outliers). We do not necessarily need this restriction, we can proceed in the entire thesis similarly with any *elliptically symmetric unimodal* distribution. Such extensions are in detail described in [2]. Formally, we can extend even to any *elliptically symmetric* distribution (but not necessary with similar results).

Definition. *A distribution with a density f is called elliptically symmetric if f takes the form*

$$f(x) = kg((x-\mu)^t \Sigma^{-1}(x-\mu)),$$

where $k \in \mathbb{R}$ is a scale factor, $g : \mathbb{R} \to \mathbb{R}$ is a non-increasing continuous function, and $\mu \in \mathbb{R}^p$ and $\Sigma \in PDS(p)$ are parameters.

Example. Elliptically symmetric distributions are the normal distribution or the multivariate t-distribution. Another example is the uniform distribution on the surface of a sphere.

We will use the term *location* for parameter μ and *scatter* for parameter Σ. Note that the scatter is not uniquely determined. If we multiply Σ with a constant we still obtain the prescription for an elliptically symmetric density function, only with different function g. Usually, this ambiguity will not be a problem. When we are in a normal case, location represents the expected value, and scatter represents the variance matrix. Having a random sample from an elliptically symmetric distribution of size n, any measurable function $L : \mathbb{R}^{p\times n} \to \mathbb{R}^p$ is called a location estimator if it does not depend on the unknown parameters. Also, any measurable function $T : \mathbb{R}^{p\times n} \to PDS(p)$ is called an estimator of scatter, if it does not depend on the unknown parameters. Our goal is to find an estimator of location and scatter with suitable properties.

1. Classical tolerance ellipsoid

Consider a bivariate normal density function and look at its level set. It is clear that it forms an ellipse. How does this ellipse look? We will show that it is the set of points $\{x \in \mathbb{R}^p : (x-\mu)^t\Sigma^{-1}(x-\mu) \leq c\}$ for a constant $c \in \mathbb{R}$. Is this really an ellipse? If so, what is the area of this ellipse? What are its semi-axes and how long are they? And how does it generalize to greater dimensions? In this section, we will answer these questions. Moreover, we will show how to estimate these ellipsoids only from a given data set.

1.1 Definition and basic properties

Definition. *For a given $\mu \in \mathbb{R}^p$ and $\Sigma \in PDS(p)$, we define Mahalanobis distance of a point $x \in \mathbb{R}^p$ as*

$$MD_{\mu,\Sigma}(x) = \sqrt{(x-\mu)^t\Sigma^{-1}(x-\mu)}.$$

In other words, the Mahalanobis distance $MD_{\mu,\Sigma}(x)$ describes how far away the point x is from the centre μ with regard to Σ. Geometrically, the graph of a function $MD_{\mu,\Sigma} : \mathbb{R}^p \to \mathbb{R}$ forms a cone in $\mathbb{R}^{p+1}$ with the peak in μ. This function can be considered as a shifted norm on $\mathbb{R}^p$, where $||x||_{\mu,\Sigma} = \sqrt{(x-\mu)^t\Sigma^{-1}(x-\mu)}$ (it is not a norm, because $||x||_{\mu,\Sigma} = 0 \iff x = 0$ is not fulfilled for $\mu \neq 0$). Now we prove a very important lemma, that helps us to understand the properties of Mahalanobis distances.

Lemma 1. *Let $Y \sim N_p(\mu, \Sigma)$ where $\Sigma \in PDS(p)$. Then*

$$(Y-\mu)^t\Sigma^{-1}(Y-\mu) \sim \chi^2_p.$$

Proof. Because Σ is positive definite, there exists $\mathbb{A} \in \mathbb{R}^{p\times p}$ non-singular, such that $\Sigma = \mathbb{A}\mathbb{A}^t$ and $Y = \mathbb{A}Z + \mu$ for $Z \sim N_p(0, \mathbb{I}_p)$ (Theorem A3). Then we can write

$$\begin{aligned}(Y-\mu)^t\Sigma^{-1}(Y-\mu) &= (\mathbb{A}Z)^t(\mathbb{A}\mathbb{A}^t)^{-1}\mathbb{A}Z = Z^t\mathbb{A}^t(\mathbb{A}^t)^{-1}\mathbb{A}^{-1}\mathbb{A}Z \\ &= Z^tZ = \sum_{k=1}^{p} Z_k^2 \sim \chi^2_p.\end{aligned}$$

□

Definition. *The 99% tolerance ellipsoid for a given $\mu \in \mathbb{R}^p$ and $\Sigma \in PDS(p)$ is defined as the set of points $x \in \mathbb{R}^p$ whose $MD_{\mu,\Sigma}(x) \leq \sqrt{\chi^2_{p,0.99}}$, in other notation $\{x \in \mathbb{R}^p : (x-\mu)^t\Sigma^{-1}(x-\mu) \leq \chi^2_{p,0.99}\}$.*

The tolerance ellipsoid is defined such that it contains 99% of the occurrence of the random vector.

Of course, usually we do not know what μ and Σ are. In classical methods, when outliers are not a concern, we can estimate them directly with the sample mean and the sample variance matrix. Therefore, for $x \in \mathbb{R}^p$ and observations $x_1, \ldots, x_n$ will *classical Mahalanobis distance* take the form $MD_{\overline{x},\mathbb{S}}(x) =$

$\sqrt{(x-\overline{x})^t\mathbb{S}^{-1}(x-\overline{x})}$, where $\overline{x} = \frac{\sum_{i=1}^n x_i}{n}$ is the sample mean and where $\mathbb{S} = \frac{\sum_{i=1}^n (x_i-\overline{x})(x_i-\overline{x})^t}{n-1}$ is the sample covariance matrix of observations $x_1, \ldots, x_n$. Also, the classical 99% tolerance ellipsoid will take the form $\{x \in \mathbb{R}^p : (x-\overline{x})^t\mathbb{S}^{-1}(x-\overline{x}) \leq \chi^2_{p,0.99}\}$.

It is important to notice that the classical tolerance ellipsoid is, in fact, an ellipsoid. That is valid only for $x_1, \ldots, x_n$ not lying in a hyperplane because only then it holds $\mathbb{S} \in PDS(p)$ (Theorem A5).

In the literature, there are several different definitions of what an ellipsoid is. One of the possible definitions is the following:

Definition. *Let $c \in \mathbb{R}^p$, and let $v_1, \ldots, v_p$ be an orthonormal basis of $\mathbb{R}^p$. Let $a_1, \ldots, a_p > 0$. Then the ellipsoid with the center c, with the semiaxes $v_1, \ldots, v_p$ of lengths $a_1, \ldots, a_p$ respectively is defined as the set*

$$E = \{[x]_V \in \mathbb{R}^p : \sum_{i=1}^p \frac{(x_i - c_i)^2}{a_i} \leq 1\}, \tag{1.1}$$

where $[x]_V = \sum_{i=1}^p x_i v_i$ is $x \in \mathbb{R}^p$ expressed in basis $V = \{v_1, \ldots, v_p\}$.

Other literature defines an ellipsoid directly as

$$E = \{x \in \mathbb{R}^p : (x-\mu)^t\Sigma^{-1}(x-\mu) \leq c\}$$

for suitable parameters. We can conclude the equivalence of these definitions from the next theorem.

Theorem 1. *The classical 99% tolerance ellipsoid for the observations $x_1, \ldots, x_n$ lying in a general position is an ellipsoid.*

Proof. Let us denote $m = \chi^2_{p,0.99} > 0$ and $c = \overline{x}$. We want to prove that $E := \{x \in \mathbb{R}^p : (x-c)^t\mathbb{S}^{-1}(x-c) \leq m\}$ is an ellipsoid from definition above.

It holds, that matrix $\mathbb{S}^{-1} \in PDS(p)$ (Theorem A5). Now, consider the SVD decomposition (Theorem A1) of $\mathbb{S}^{-1} = \mathbb{U}\mathbb{D}\mathbb{U}^t$ for orthogonal $\mathbb{U} \in \mathbb{R}^{p\times p}$ and diagonal $\mathbb{D} \in \mathbb{R}^{p\times p}$ with $\mathbb{D}_{i,i} > 0$ elements on the diagonal. Then

$$\begin{aligned} E &= \{x \in \mathbb{R}^p : (x-c)^t\mathbb{U}\mathbb{D}\mathbb{U}^t(x-c) \leq m\} \\ &= \{x \in \mathbb{R}^p : (\mathbb{U}^t(x-c))^t\mathbb{D}(\mathbb{U}^t(x-c)) \leq m\} \\ &= \{[x]_{U^t} \in \mathbb{R}^p : (x-c)^t\mathbb{D}(x-c) \leq m\} \\ &= \{[x]_{U^{-1}} \in \mathbb{R}^p : \sum_{i=i}^p \frac{(x_i-c_i)^2}{\frac{m}{\mathbb{D}_{i,i}}} \leq 1\}, \end{aligned}$$

what is actually the definition of an ellipsoid with the basis U^{-1}. Note, that it is a set of the eigenvectors of $\mathbb{S}^{-1}$.

Now, we can find the shape of this ellipsoid. It has its centre in c. If we denote $e_1, \ldots, e_p$ the standard basis of $\mathbb{R}^p$, then its principal semi-axes are $\mathbb{U}^{-1}(e_1), \ldots, \mathbb{U}^{-1}(e_p)$, with lengths $\frac{m}{\mathbb{D}_{1,1}}, \ldots, \frac{m}{\mathbb{D}_{p,p}}$ respectively. Visualization is given on Figure 1.1. □

1.2 Volume of an ellipsoid

In this section we will find the volume of the classical tolerance ellipsoid, generally given by

$$E = \{x \in \mathbb{R}^p : (x-t)^t \mathbb{C}^{-1}(x-t) \leq c_1\}, \tag{1.2}$$

for some $\mathbb{C} \in PDS(p)$, $t \in \mathbb{R}^p$ and $c_1 > 0$. We will use this result in Chapter 4.

Lemma 2. *The volume of a p-dimensional ellipsoid E with lengths of its semi-axes $a_1, a_2, \dots, a_p$ is given by*

$$V(E) = \frac{2}{p}\frac{\pi^{p/2}}{\Gamma(p/2)}\prod_{i=1}^{p} a_i.$$

Proof. It is an easy application of integration using spherical or ellipsoidal coordinates. See [3]. □

Theorem 2. *Let $\mathbb{C} \in PDS(p)$, $t \in \mathbb{R}^p, c_1 > 0$. The volume of the p-dimensional ellipsoid from* (1.2) *is given by*

$$V(E) = \frac{2}{p}\frac{\pi^{p/2}}{\Gamma(p/2)} c_1^p det(\mathbb{C}).$$

Proof. We aim to prove, that the product of the lengths of the principal semiaxes is equal to $c_1^p det(\mathbb{C})$. Then the proof is complete due to Lemma 2. Let us use the SVD decomposition (Theorem A1) in the form $\mathbb{C}^{-1} = \mathbb{U}\mathbb{D}\mathbb{U}^t$ for diagonal $\mathbb{D}$ with $\mathbb{D}_{i,i}$ on the diagonal, and $\mathbb{U}$ an orthogonal matrix. In the proof of Theorem 1 we showed that E is an ellipsoid with principal semiaxes with lengths $\frac{c_1}{\mathbb{D}_{1,1}}, \dots, \frac{c_1}{\mathbb{D}_{p,p}}$. But it is well known that in the SVD decomposition $\mathbb{D}_{i,i}$ are the eigenvalues of $\mathbb{C}^{-1}$. Now we use that a determinant of a matrix is the product of its eigenvalues. Therefore, the product of the length of the principal semi axes of E is equal to $\prod_{i=1}^{p} \frac{c_1}{\mathbb{D}_{i,i}} = \frac{c_1^p}{\prod_{i=1}^{p} \mathbb{D}_{i,i}} = \frac{c_1^p}{det(\mathbb{C}^{-1})} = c_1^p det(\mathbb{C})$ what we wanted to prove. □

Remark. *Let $c_1 > 0$. For every $\mathbb{C} \in PDS(p)$, $t \in \mathbb{R}^p$ there exists a uniquely determined ellipsoid with the form* (1.2). *This implication can be reversed. For every ellipsoid there exists a uniquely determined $\mathbb{C} \in PDS(p)$, $t \in \mathbb{R}^p$ such that this ellipsoid has the form from* (1.2). *Therefore, estimating the location and scatter of an elliptically symmetric distribution is equivalent with finding an ellipsoid which is a level set of the density function.*

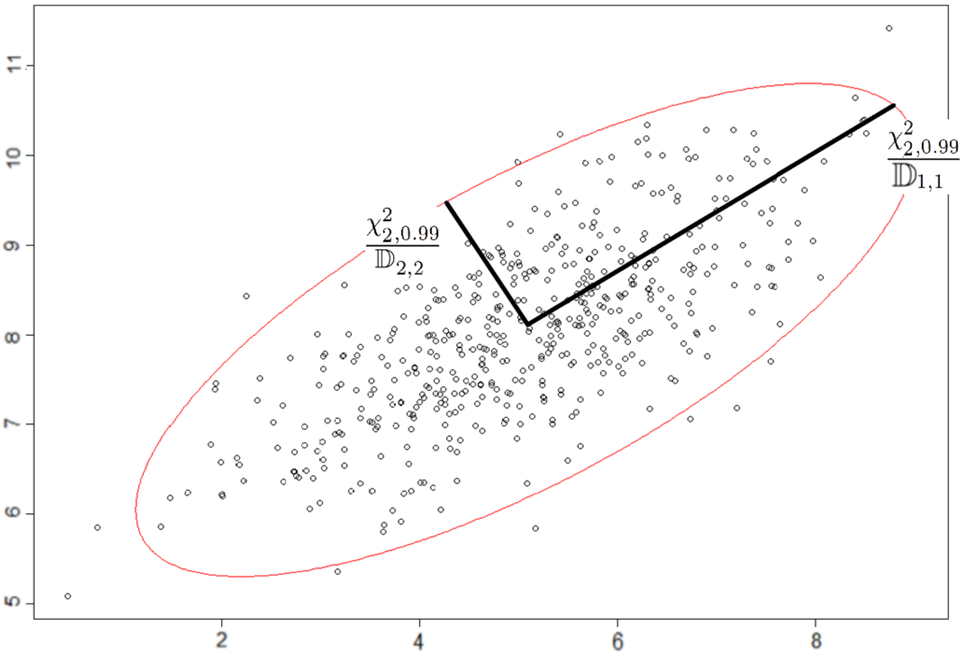


Figure 1.1: We concluded that the 99% tolerance ellipsoid computed from $x_1, \ldots, x_n$ will have its centre in $\overline{x}$ and semi axes $\mathbb{U}^{-1}(e_1), \ldots, \mathbb{U}^{-1}(e_p)$ with lengths $\frac{\chi^2_{p,0.99}}{\mathbb{D}_{1,1}}, \ldots, \frac{\chi^2_{p,0.99}}{\mathbb{D}_{p,p}}$ (where $\mathbb{S}^{-1} = \mathbb{UDU}^t$ is the SVD decomposition, $\mathbb{D}_{i,i}$ is the i-th diagonal element of $\mathbb{D}$ and $e_1, \ldots, e_p$ is the standard basis of $\mathbb{R}^p$). We also concluded that the volume of this ellipsoid equals $K_p det(\mathbb{S})$ for some constant K_p depending only on the dimension p.

2. MCD estimator

One of the most commonly used robust estimators is the MCD estimator of location and scatter.

The main idea of the MCD estimator is, that we look only at some of the observations and compute the location and scatter only from them. But which $h \leq n$ observations should we choose? As long as we are in an elliptically symmetric distribution, the most suitable observations are those that are closest to each other, because they tend to be in the centre of the distribution. Roughly speaking, if there are many of them close together, most likely they will not be error measurements. The term "close" can be mathematically interpreted in many ways, each estimator then takes a different interpretation.

The MCD estimator considers the points with the minimal determinant of the sample covariance matrix. As we can see in Figure 2.1, if we take only a fraction of our dataset, points in the middle tend to have a smaller determinant of the sample covariance matrix (represented by the volume of the ellipses). From the fraction of points in the centre of the distribution, we can compute the sample covariance matrix. We will show, that by multiplying this sample covariance matrix by a suitable constant, we will obtain a consistent estimator of Σ (for normal distributions).

We can now move on to the formal definition of the MCD estimator.

Definition. *The raw MCD estimator with parameter $h \in \mathbb{N}$, where $n/2 + 1 \leq h \leq n$, defines the mean and covariance matrix as follows:*

$\widehat{\mu}_{MCD}$ *is the mean of those h observations whose determinant of the sample covariance matrix is minimal.*

$\widehat{\Sigma}_{MCD}$ *is the corresponding covariance matrix multiplied by a consistency factor $c_0 = q/F_{\chi^2_{p+2}}(\chi^2_{p,q})$ where $q = h/n$ and $F_{\chi^2_{p+2}}$ is the distribution function of the chi-squared distribution with $p + 2$ degrees of freedom.*

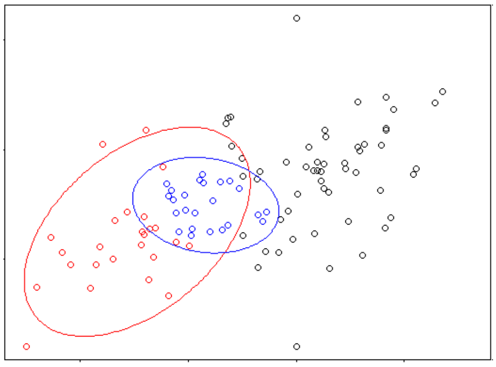


Figure 2.1: Points in the middle of the distribution tend to be more closely packed. Therefore, they have smaller determinant of the covariance matrix and are more suitable for our estimator. In the figure there is the same number of red and blue points. The ellipses visualise the determinants of covariance matrices computed from the blue and red points. An ellipse with smaller area will have smaller determinant.

The consistency factor c_0 is chosen in order to obtain consistency at the normal distribution. In other words, we inflate the covariance matrix, such that we would expect a covariance matrix for a normal distribution to behave when we have only those h observations. We discuss the consistency factor in Section 3.2.2.

Now, we have a different, robust tolerance ellipsoid based on the MCD estimator.

Definition. *For $x \in \mathbb{R}^p$ we define the robust Mahalanobis distance as*

$$RD(x) = \sqrt{(x - \widehat{\mu}_{MCD})^t \widehat{\Sigma}_{MCD}^{-1} (x - \widehat{\mu}_{MCD})}$$

and the robust tolerance 99 % ellipse based on the MCD estimator are those $x \in \mathbb{R}^p$ that satisfy $RD(x) \leq \sqrt{\chi^2_p(0.99)}$.

Note that MCD can be computed only if $h > p$, otherwise the covariance matrix of any h-subset will be singular. It is recommended that $n > 5p$ [4].

2.1 Applications of MCD

As an example, consider a dataset on waste material in Slovak boroughs (dataset is available in [5]). Each borough is obliged to monitor the quantities of all types of waste material such as glass waste, paper waste and many other. For each ton of waste pays the borough a fee. In Slovakia, there are 81 boroughs, and we will focus on the 10 most common types of waste materials, therefore we have $n = 81$ observations with dimension $p = 10$.

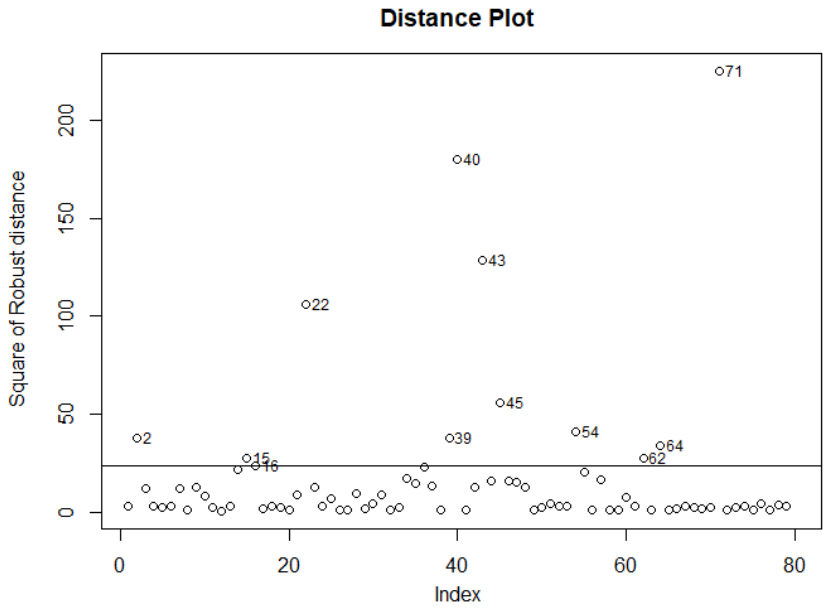


Figure 2.2: Robust Mahalanobis distances of the observations of the waste quantity in Slovak boroughs. Each observation represents one Slovak borough (for example, borough with index 1 is Bratislava I.). Boroughs above the horizontal cut-off line (above $\chi^2_{10}(0.99) = 23.2$) are marked as outliers. We can conclude that the borough with index 71 (Košice II) deviates the most.

In our model, we can assume that the quantity of waste for one resident in each borough follows a normal distribution (we compute the waste material only per capita, otherwise there would be much more waste material in larger boroughs). For simplicity, we also assume that the quantity of the waste material is in each borough independent (which is not exactly true, e.g. if there is a factory for plastic in a nearby borough, it is expected to be a more plastic waste).

Our goal is to investigate which boroughs are not monitoring their waste material as expected, or where is remarkably more waste materials used. We will use the robust distances based on the MCD estimates to mark outliers. In Figure 2.2 we have data visualized with a distance plot so we can see which boroughs deviate the most. For example, the borough with index 71 is Košice II. This makes sense because the large factory (US Steel Košice) is based here.

2.2 MCD algorithm

Computing the MCD estimator exactly is computationally expensive, because it requires as many as $\binom{n}{h}$ evaluations. Therefore, one usually resorts to finding only an approximate solution. We can use the FAST-MCD algorithm [6] that uses iteration and repetition. The main part of the algorithm is this:

1. Take a random h-subset of $(x_1, ..., x_n)$ and compute $\overline{x}_h, \mathbb{S}_h$ the sample mean and covariance matrix of this subset;

2. For each $i \leq n$ compute the relative distances $D_i := MD_{\overline{x}_h, \mathbb{S}_h}(x_i)$;

3. Take a new h-subset of $(x_1, ..., x_n)$ that consists of those elements with minimal relative distances D_i. Compute $\overline{x}^*_h, \mathbb{S}^*_h$ (the sample mean and the sample matrix from this subset). If $det(\mathbb{S}^*_h) < det(\mathbb{S}_h)$ then set $\overline{x}_h := \overline{x}^*_h$, $\mathbb{S}_h := \mathbb{S}^*_h$ and go to step 2. If $det(\mathbb{S}^*_h) = det(\mathbb{S}_h)$ then end the algorithm with an output $\overline{x}^*_h, \mathbb{S}^*_h$.

Lemma 3. *Denote $\mathbb{S}_h, \mathbb{S}^*_h$ as in the previous algorithm. Then it is always satisfied that $det(\mathbb{S}^*_h) \leq det(\mathbb{S}_h)$.*

Proof. The proof is relatively long and technical, and can be found in [6]. □

This algorithm ends in a finite number of steps because there is only a finite number of h-subsets of $(x_1, ..., x_n)$. Note that it generally does not return the global minimum of $det(\mathbb{S}_h)$.
One of the possible approaches is to repeat this algorithm many times with new random initial subsets (in the package *"robustbase v0.93-4"* has this algorithm in the programming language `R` this default value 30000 repetitions) and finally to return a solution which can be expected to be close to the minimal solution.

There are several improvements of this algorithm that can be found in the literature. For example, we can repeat step 2 only twice, because the resulting determinant is usually sufficiently close to the determinant obtained from further iterations. The time complexity of this algorithm is $O(p^3 n \log n)$ with a large constant. More sophisticated methods using Cholesky decomposition, along with a discussion on some numerical properties of this algorithm, can be found in [7].

3. Properties of MCD estimator

In this chapter, we aim to prove some important properties of the MCD estimator. Is it robust, and if so, how robust? How many outliers can this estimator handle? Does it really estimate Σ? More formally, is $\widehat{\Sigma}_{MCD}$ a consistent estimator of scatter? At first, we will recall some definitions that are important in robust statistics, that we want to examine.

We will refer to the consistency as the weak consistency (convergence in probability). Again, let $X_1, \ldots, X_n$ be a random sample from an elliptically symmetric distribution with parameter of location μ and scatter Σ. Suppose we have data from this distribution, the i-th observation is denoted by $x_i = (x_{i,1}, ...x_{i,p})^t$, and stored in columns of the matrix $\mathbb{M} = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$. For convenience, we will use the notation where matrix $\mathbb{M}$ contains fixed observations, not random variables.

3.1 General definitions

Definition. *Let $L : \mathbb{R}^{p\times n} \to \mathbb{R}^p$ be a location estimator, $T : \mathbb{R}^{p\times n} \to PDS(p)$ be an estimator of scatter. An affine equivariant estimator of location and scatter $(L(\mathbb{M}), T(\mathbb{M}))$ is one for which:*

- $L(\mathbb{A}\mathbb{M} + b) = \mathbb{A}L(\mathbb{M}) + b$,
- $T(\mathbb{A}\mathbb{M} + b) = \mathbb{A}T(\mathbb{M})\mathbb{A}^t$,

for any non-singular matrix $\mathbb{A} \in \mathbb{R}^{p\times p}$ and $b \in \mathbb{R}^p$. Here, we write $\mathbb{A} + b$ the vector b added to every column of the matrix $\mathbb{A}$.

Basically, affine equivariance means that an estimator is well adjusted for affine transformations of the data. Affine equivariance is an important property because an affine equivariant estimator remains consistent after an affine transformation.

3.1.1 Robustness

One of the most common definitions describing the robustness of an estimator is the breakdown point. The breakdown point represents the minimal percentage of observations that can carry our estimate beyond all bounds. In other words, the number of points that can do with the estimate whatever they want, if they are suitably chosen. For example, the sample mean and the sample variance have the breakdown point $1/n$, because a single element can change those estimates arbitrarily.

Definition. *Let $(L(\mathbb{M}), T(\mathbb{M}))$ be a location and scatter estimator, and let us denote by $\mathbb{M}^{(k)}$ the set of all matrices obtained by replacing k columns (k data vectors) in $\mathbb{M}$ with arbitrary points. For the location estimator we define a breakdown point as follows:*

$$BP(L, \mathbb{M}) := \frac{1}{n}(min\{k \in \{1, ..., n\} : \sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} ||L(\mathbb{M}) - L(\mathbb{M}_k)|| = \infty\}).$$

For the scatter estimator we define a breakdown point as follows:

$$BP(T,\mathbb{M}) := \frac{1}{n}(min\{k \in \{1,...,n\} : \\ \sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} \max_i |log(\lambda_i(T(\mathbb{M}))) - log(\lambda_i(T(\mathbb{M}_k)))| = \infty\}),$$

where $\lambda_i(\mathbb{A})$ denotes i-th greatest eigenvalue of $\mathbb{A}$.

The transcription for scatter means that arbitrary points can carry some eigenvalue arbitrarily close to 0 or beyond all bounds; using logarithms only expresses these two options.

We will show one important theorem, which describes the upper bound for the breakdown point of an affine equivariant estimator.

Theorem 3. *Let $(L(\mathbb{M}), T(\mathbb{M}))$ be a location and scatter estimator, that is affine equivariant. Let $b \in \mathbb{R}^p$ and $\mathbb{A} \in \mathbb{R}^{p\times p}$ be nonsingular matrix. Then it holds*

1. $BP(L,\mathbb{M}) = BP(L, \mathbb{A}\mathbb{M} + b)$;
2. $BP(L,\mathbb{M}) \leq \frac{\lfloor \frac{n+1}{2} \rfloor}{n}$;
3. $BP(T,\mathbb{M}) = BP(T, \mathbb{A}\mathbb{M} + b)$;
4. $BP(T,\mathbb{M}) \leq \frac{\lfloor \frac{n-p+1}{2} \rfloor}{n}$, *as long as $x_1, \ldots, x_n$ lie in a general position.*

Proof. We will prove only the first two statements. The remaining two statements can be proven similarly or can be also found in [8].

(1): Let $b \in \mathbb{R}^p$ and $\mathbb{A} \in \mathbb{R}^{p\times p}$ be a nonsingular matrix. Let $k = nBP(L,\mathbb{M})$, therefore $k \in \{1,...,n\}$ and is the minimal k for which $\sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} ||L(\mathbb{M}) - L(\mathbb{M}_k)|| = \infty$.

" $\leq$ ": Let $\mathbb{M}^{(k-1)}$ be a set of matrices obtained by replacing $k-1$ columns in $\mathbb{M}$ with arbitrary points. Then it holds (using Theorem A4)

$$\begin{aligned} &\sup_{\mathbb{M}_{k-1} \in \mathbb{M}^{(k-1)}} ||L(\mathbb{A}\mathbb{M} + b) - L(\mathbb{A}\mathbb{M}_{k-1} + b)|| \\ &= \sup_{\mathbb{M}_{k-1} \in \mathbb{M}^{(k-1)}} ||\mathbb{A}L(\mathbb{M}) + b - \mathbb{A}L(\mathbb{M}_{k-1}) - b|| \\ &= \sup_{\mathbb{M}_{k-1} \in \mathbb{M}^{(k-1)}} ||\mathbb{A}(L(\mathbb{M}) - L(\mathbb{M}_{k-1}))|| \\ &\leq ||\mathbb{A}|| \sup_{\mathbb{M}_{k-1} \in \mathbb{M}^{(k-1)}} ||L(\mathbb{M}) - L(\mathbb{M}_{k-1})|| < \infty. \end{aligned}$$

Therefore $BP(L, \mathbb{A}\mathbb{M} + b) > \frac{k-1}{n}$.

" $\geq$ ": Let λ be the smallest eigenvalue of $\mathbb{A}$. Then it holds (using Theorem A4)

$$\sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} ||L(\mathbb{A}\mathbb{M} + b) - L(\mathbb{A}\mathbb{M}_k + b)|| = \sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} ||\mathbb{A}L(\mathbb{M}) + b - \mathbb{A}L(\mathbb{M}_k) - b|| \\ = \sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} ||\mathbb{A}(L(\mathbb{M}) - L(\mathbb{M}_k))|| \geq |\lambda| \sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} ||L(\mathbb{M}) - L(\mathbb{M}_k)|| = \infty.$$

Therefore $BP(L, \mathbb{AM} + b) \leq \frac{k}{n}$.

(2): Let $k = \lfloor \frac{n+1}{2} \rfloor$. For a contradiction, let us assume that

$$\sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} ||L(\mathbb{M}) - L(\mathbb{M}_k)|| < \infty.$$

This implies that there exists $Q \in \mathbb{R}$ such that $||L(\mathbb{M}_k)|| < Q$ for all $\mathbb{M}_k \in \mathbb{M}^{(k)}$. Because $k \geq n/2$, $\mathbb{M}_k$ will have changed at least half of the columns of $\mathbb{M}$. Let $v \in \mathbb{R}^p$ and let $\mathbb{M}_k$ be the matrix with columns $(x_1, \ldots, x_{\lfloor n/2 \rfloor}, x_{\lfloor n/2 \rfloor + 1} + v, \ldots, x_n + v)$. In words, we shifted the second half of the columns. It is important to notice, that $\mathbb{M}_k \in \mathbb{M}^{(k)}$. But also $(\mathbb{M}_k - v) \in \mathbb{M}^{(k)}$, therefore it holds $Q > ||L(\mathbb{M}_k - v)||$. Now, because L is affine equivariant, it holds that $||L(\mathbb{M}_k - v)|| = ||L(\mathbb{M}_k) - v||$. But for v large enough, both conditions $Q > ||L(\mathbb{M}_k) - v||$ and $Q > ||L(\mathbb{M}_k)||$ cannot be satisfied, which is a contradiction. □

3.1.2 Efficiency

We want to know how accurate is our estimator. We can compare two estimators by their variances, the smaller variance of estimation the better. It is easier to understand how “good” the variance is, if we divide it with the “best” variance that can be achieved. If our estimator is unbiased, this property can be expressed as the *efficiency* of our estimator [9]. In a normal distribution, the sample mean is an *efficient* estimator of the expected value, which means that the variability of this estimator is the smallest possible (out of all unbiased estimators).

Often, we do not know if our estimator is unbiased. Anyway, if we know the efficient estimator, we can still denote by *efficiency* the quotient of the variance of our estimator and the variance of the efficient estimator. We want our estimate to have the efficiency closest to one. Note that by this definition, we can also obtain the efficiency greater than one.

For a multivariate case, the efficiency cannot be defined as the quotient of the variances, because quotient of matrices is not defined. Instead, we can look only at the diagonal elements of the matrix and we define multivariate efficiency as the smallest efficiency among them. Therefore, we will take the “worst” of the diagonal elements with which we proceed as in the univariate case. Some other definitions are often used, such as the average of efficiencies of all elements of the matrix.

3.2 Properties of the MCD estimator

3.2.1 Affine equivariance

Theorem 4. *MCD estimator is an affine equivariant estimator.*

Proof. Let $\mathbb{A} \in \mathbb{R}^{p \times p}$ be non-singular and $b \in \mathbb{R}^p$. We want to prove that

- $\widehat{\mu}_{MCD}(\mathbb{AM} + b) = \mathbb{A}\widehat{\mu}_{MCD} + b$,
- $\widehat{\Sigma}_{MCD}(\mathbb{AM} + b) = \mathbb{A}\widehat{\Sigma}_{MCD}\mathbb{A}^t$.

If we denote $H \subset \{1, ..., n\}$ any set with h elements (for $h > p$), $\mathbb{M}_H \in \mathbb{R}^{p \times h}$ the matrix obtained by erasing every column (observation) of $\mathbb{M}$ whose index is not in H, and $S : \mathbb{R}^{p \times h} \to \mathbb{R}^{p \times p} : \mathbb{M}_H \mapsto S(\mathbb{M}_H)$ the sample covariance matrix computed from the columns of matrix $\mathbb{M}_H$, then it holds

$$det(S(\mathbb{A}\mathbb{M}_H + b)) = det(\mathbb{A}S(\mathbb{M}_H)\mathbb{A}^t) = det(\mathbb{A})^2 det(S(\mathbb{M}_H)).$$

The first equality holds due to

$$\begin{aligned} S(\mathbb{A}\mathbb{M}_H + b) &= \frac{1}{h-1}\sum_{i=1}^{h}(\mathbb{A}x_{H_i} - \mathbb{A}\overline{x}_H)(\mathbb{A}x_{H_i} - \mathbb{A}\overline{x}_H)^t \\ &= \frac{1}{h-1}\sum_{i=1}^{h}\mathbb{A}(x_{H_i} - \overline{x}_H)(x_{H_i} - \overline{x}_H)^t\mathbb{A}^t = \mathbb{A}S(\mathbb{M}_H)\mathbb{A}^t, \end{aligned}$$

where x_{H_i} is i-th column of $\mathbb{M}_H$ and $\overline{x}_H = \frac{1}{h}\sum_{i=1}^{h} x_{H_i}$. So $\mathbb{M}_H$ minimizes the covariance determinant (with respect to the untransformed data) if and only if $\mathbb{A}\mathbb{M}_H$ minimizes the covariance determinant (with respect to the transformed data). Therefore, $\widehat{\mu}_{MCD}$ and $\widehat{\Sigma}_{MCD}$ will be computed from the same h points (before and after the affine transformation). Because the sample mean and the sample covariance matrix are affine equivariant estimators [8], and $\widehat{\mu}_{MCD}$, $\widehat{\Sigma}_{MCD}$ are computed as the sample mean and sample covariance matrix of h observations, therefore $\widehat{\mu}_{MCD}$ and $\widehat{\Sigma}_{MCD}$ are also affine equivariant estimators. □

3.2.2 Consistency factor

We will show that after choosing $c_0 = q/F_{\chi^2_{p+2}}(\chi^2_{p,q})$, we obtain consistency for normal distributions. We will show only a main idea of the proof because one step appears to be quite difficult, although intuitive. We assume $p = 1$, for $p > 1$ we can proceed very similarly.

If we want to compute the consistency factor, it is sufficient to compute this only for $X_1, \ldots, X_n \sim N(0, 1)$. Due to the affine equivariance of the MCD estimator, consistency remains the same after any affine transformation, therefore for any $N(\mu, \sigma^2)$. For $N(0, 1)$, it is intuitive that (asymptotically for $n \to \infty$) the $\widehat{\Sigma}_{MCD}$ will be computed from the "middle" $q\,\%$ of the standard normal distribution (the meaning of this expression is that $\widehat{\Sigma}_{MCD}$ will be the same as the variance of such a truncated standard normal distribution).

Definition. *Let $\mu \in \mathbb{R}$, $\sigma > 0$. We define the truncated normal distribution with parameter $y > 0$ to be the distribution with density f, for which $f(x) = 0$ for $|x| > y$, and for $x \leq y$ the function f is defined as the density of the normal distribution $N(\mu, \sigma^2)$ multiplied by an appropriate constant. We will denote this distribution by $N^y(\mu, \sigma^2)$.*

Density function is shown in Figure 3.1.

Theorem 5. *Let $q \in (0, 1)$. Denote by $\phi_1(x)$ the density of the standard normal distribution. Let $N^y(0, 1)$ be the truncated normal distribution for $y > 0$, such that $\int_{-y}^{y} \phi_1(x)dx = q$ (this refers to middle $q\,\%$ of the standard normal distribution). Then, the variance of $N^y(\mu, \sigma^2)$ is equal to $\frac{F_{\chi^2_{1+2}}(\chi^2_{1,q})}{q}$.*

Proof. We will use one short auxiliary lemma:

Lemma. *Let u_t be the t-quantile of the standard normal distribution. Then for every $t \in (\frac{1}{2}, 1)$ holds $u_t = \chi_{1,(2t-1)}$ (where $\chi_{p,q}$ refers to the square root of the q-quantile of chi-squared distribution with p degrees of freedom).*

Proof. Let $Z \sim N(0,1)$. Then $t = P(Z \leq u_t) = P(Z \in (0, u_t)) + \frac{1}{2} = \frac{P(Z \in (-u_t, u_t))}{2} + \frac{1}{2} = \frac{P(Z^2 \leq u_t^2)}{2} + \frac{1}{2}$. We can conclude that $2t - 1 = P(Z^2 \leq u_t^2)$. Because Z^2 has chi-squared distribution with one degree of freedom, we obtain $\chi^2_{1,(2t-1)} = u_t^2$. □

Now, we can calculate the restriction boundaries $1_{(-y,y)}$ by putting into equation

$$\int_{-y}^{y} \phi_1(x)dx = q,$$

where we easily compute $y = u_\delta$ for $\delta = \frac{q+1}{2}$. Now we have a truncated standard normal distribution, whose density will be ϕ_1 divided by q (to obtain $\int_{-u_\delta}^{u_\delta} \frac{\phi_1(x)}{q}dx = 1$), which we denote by $\widetilde{\phi}_1(x) := \frac{\phi_1(x)}{q} 1_{(-u_\delta, u_\delta)}(x)$.

The variance of this truncated distribution is equal to

$$\int_{-\infty}^{\infty} x^2 \widetilde{\phi}_1(x)dx = \int_{-u_\delta}^{u_\delta} x^2 \frac{\phi_1(x)}{q} dx = \int_{-\chi_{1,q}}^{\chi_{1,q}} x^2 \frac{1}{q\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

$$= \frac{\sqrt{2}}{q\sqrt{\pi}} \int_0^{\chi_{1,q}} x^2 e^{-\frac{x^2}{2}} dx = \frac{2}{q\sqrt{\pi}} \int_0^{\frac{\chi^2_{1,q}}{2}} t^{1/2} e^{-t} dt = \frac{\gamma(3/2, \frac{\chi^2_{1,q}}{2})}{\Gamma(3/2)} \frac{1}{q} = \frac{F_{\chi^2}(\chi^2_{1,q})}{q}.$$

The second equality holds due to the previous lemma, the third one due to symmetry. At the fourth we used the substitution $t = \frac{x^2}{2}$. The fifth equation holds by definition $\gamma(k, x) = \int_0^x t^{k-1} e^{-t} dt$, and because $\Gamma(3/2) = \sqrt{\pi}/2$. The last equation is valid due to the definition of the chi-squared distribution [10]. □

If it holds that $\widehat{\Sigma}_{MCD}$ converges to $\frac{F_{\chi^2_{1+2}}(\chi^2_{1,q})}{q}$, then we need to multiply our estimate by $\frac{q}{F_{\chi^2_{1+2}}(\chi^2_{1,q})}$ to obtain a consistent estimate for $\sigma = 1$. The difficult step (that $\widehat{\Sigma}_{MCD}$ really converges) uses more sophisticated mathematics; parts of this proof can be found in [11] and [12].

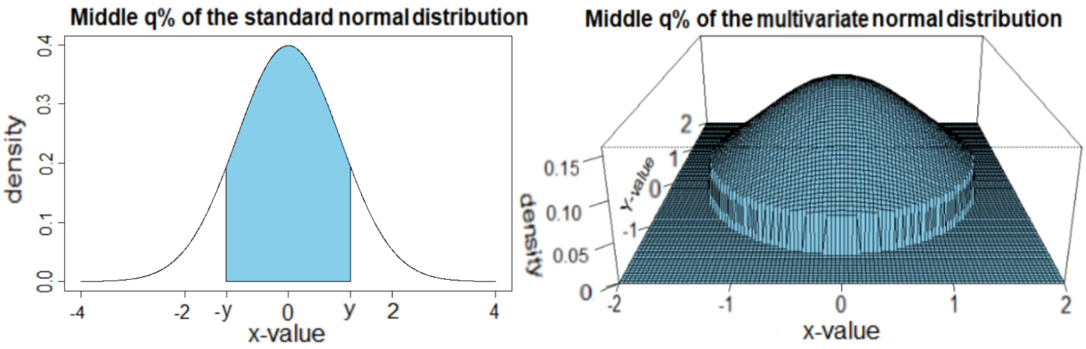


Figure 3.1: On the left panel is the density of a truncated normal distribution $N^y(0,1)$. On the right panel is a possible generalization of the density of a truncated normal distribution for $p = 2$.

3.2.3 Robustness

We will prove that the MCD estimator has the largest possible breakdown point, that can be attained for an affine equivariant estimator. First, we will need two auxiliary lemmas.

Lemma 4. *Let $n > p$ and let $x_1, \dots, x_n \in \mathbb{R}^p$ be in a general position (no more than p points lie on any hyperplane of dimension less than p). Then it holds that $\mathbb{S}_c := \frac{\sum_{k=1}^n (x_k - c)(x_k - c)^t}{n-1} \in PDS(p)$ for every $c \in \mathbb{R}^p$. Moreover, it holds that $\lambda_p(\mathbb{S}_c) \geq \lambda_p(\mathbb{S}_{\overline{x}_n})$, where $\lambda_p(\mathbb{S}_c)$ denotes the smallest eigenvalue of $\mathbb{S}_c$.*

Proof. For non-zero $y \in \mathbb{R}^p$ is

$$y^t \mathbb{S}_c y = \frac{1}{n-1} \sum_{k=1}^n y^t (x_k - c)(x_k - c)^t y = \frac{1}{n-1} \sum_{k=1}^n ((x_k - c)^t y)^2 \geq 0.$$

It can be zero if and only if $(x_k - c)^t y = 0$ for all k. But because $(x_k - c)$ spans $\mathbb{R}^p$, there exist $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ such that $y = \sum_{i=1}^n \alpha_i (x_i - c)$. This gives us $y^t y = \sum_{i=1}^n \alpha_i (x_i - c)^t y$. Therefore if $(x_k - c)^t y = 0$ for all $k \leq n$, then $y = 0$, which is a contradiction. Therefore $y^t \mathbb{S}_c y > 0$.

Concerning the statement $\lambda_p(\mathbb{S}_c) \geq \lambda_p(\mathbb{S}_{\overline{x}_n})$, the smallest eigenvalue in $\mathbb{S}_c$ is equal to

$$\begin{aligned} \lambda_p(\mathbb{S}_c) &= \min_{y \in \mathbb{R}^p : ||y||=1} y^t \mathbb{S}_c y = \min_{y \in \mathbb{R}^p : ||y||=1} \frac{1}{n-1} \sum_{k=1}^n ((x_k - c)^t y)^2 \\ &\geq \min_{y \in \mathbb{R}^p : ||y||=1} \frac{1}{n-1} \sum_{k=1}^n ((x_k - \overline{x}_n)^t y)^2 = \min_{y \in \mathbb{R}^p : ||y||=1} y^t \mathbb{S}_{\overline{x}} y = \lambda_p(\mathbb{S}_{\overline{x}_n}). \end{aligned}$$

The inequality holds, because of Theorem A5 (the sample mean minimizes the squared error). The first equality (and the last one) holds because: for the SVD decomposition (Theorem A1) in the form $\mathbb{S}_c = \mathbb{U}^t \mathbb{D} \mathbb{U}$, with orthogonal $\mathbb{U}$ and diagonal $\mathbb{D}$ with $\mathbb{D}_{i,i} = \lambda_i(\mathbb{S}_c)$, it holds that

$$\begin{aligned} \lambda_p(\mathbb{S}_c) &= \min_{y \in \mathbb{R}^p : ||y||=1} y^t \mathbb{D} y = \min_{y \in \mathbb{R}^p : ||y||=1} (\mathbb{U}y)^t \mathbb{D} (\mathbb{U}y) \\ &= \min_{y \in \mathbb{R}^p : ||y||=1} y^t \mathbb{U}^t \mathbb{D} \mathbb{U} y = \min_{y \in \mathbb{R}^p : ||y||=1} y^t \mathbb{S}_c y. \end{aligned}$$

□

Lemma 5. *Let $x_1, \dots, x_{p+1} \in \mathbb{R}^p$ be in a general position. Then there exists $Q > 0$ such that for all $x_{p+2}, \dots, x_n \in \mathbb{R}^p$ is every eigenvalue of $\mathbb{S}_n$ greater than Q. Here, $\mathbb{S}_n$ stands for the sample covariance matrix computed from all $x_1, \dots, x_n$.*

Proof. It is sufficient to prove this only for one arbitrary point ($n = p + 2$), otherwise we can repeat the following step $n - p - 1$ times. WLOG $\overline{x}_n = 0$, otherwise we shift our points (this does not change the sample covariance matrix).

Let $\mathbb{S}^*_{p+1} = \frac{1}{p+1} \sum_{i=1}^{p+1} (x_i - \overline{x}_n)(x_i - \overline{x}_n)^t = \frac{1}{p+1} \sum_{i=1}^{p+1} x_i x_i^t$. We know (Lemma 4) that $\mathbb{S}^*_{p+1} \in PDS(p)$, and denote by Q_1 its smallest eigenvalue ($Q_1 > 0$). Note, that $Q_1 \geq Q > 0$, where Q is the smallest eigenvalue of $\frac{1}{p+1} \sum_{i=1}^{p+1} (x_i - \overline{x}_{p+1})(x_i -$

$\overline{x}_{p+1})^t$ (this is due the second part of Lemma 4), which does not depend on $x_{p+2}, \ldots, x_n$.

Now, we know that $\mathbb{S}_n = \mathbb{S}^*_{p+1} + \frac{1}{p+1} x_n x_n^t$. From the definition of positive definite matrices, it holds that $\forall h \in \mathbb{R}^p, ||h|| = 1 : h^t \mathbb{S}^*_{p+1} h \geq Q_1 > 0$ and $h^t(\frac{1}{p+1} x_n x_n^t)h = \frac{1}{p+1}(h^t x_n)(h^t x_n)^t \geq 0$. Together, we must have $h^t \mathbb{S}_n h = h^t(\mathbb{S}^*_{p+1} + \frac{1}{p+1} x_n x_n^t)h \geq h^t \mathbb{S}^*_{p+1} h \geq Q_1 \geq Q > 0$. Therefore, the smallest eigenvalue of $\mathbb{S}_n$ is greater than or equal to Q. □

Theorem 6. *The breakdown point of the MCD estimator of scatter with parameter h is equal to $\frac{1}{n} min\{n-h+1, h-p\}$. It is maximal for $h = \left\lceil \frac{n+p+1}{2} \right\rceil$, when the breakdown point is equal to $BP(\widehat{\Sigma}_{MCD}, \mathbb{M}) = \frac{\lfloor \frac{n-p+1}{2} \rfloor}{n}$. This is the largest possible breakdown point, that can be attained for an affine equivariant estimator.*

Proof. We will prove that the breakdown point is equal to $\frac{1}{n} min\{n-h+1, h-p\}$. This value is indeed maximal for $h = \left\lceil \frac{n+p+1}{2} \right\rceil$. We proved in Theorem 3 that we cannot hope for a better result than this. Let $h - p \geq n - h + 1$ (in other form $h \geq \frac{n+p+1}{2}$), otherwise we proceed similarly.

Roughly speaking, we want to find out how many "bad" points can be replaced in $\mathbb{M}$ so that the estimate will not go beyond all bounds. Therefore, we say that matrix $\mathbb{M}$ contains "good" points, which means that it contains a given $x_1, \ldots, x_n$ lying in a general position.

" $\leq$ ": Put $k = n-h+1$. We want to show $\sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} \max_i |log(\lambda_i(T(\mathbb{M}))) - log(\lambda_i(T(\mathbb{M}_k)))| = \infty$, or that after replacing $n-h+1$ points with arbitrary points, the estimate can be carried beyond all bounds. This is trivial, if we compute the MCD estimate from h points, at least one of them will be an arbitrary one, therefore it can carry the estimate beyond all bounds (already a single arbitrary element can carry the sample variance beyond all bounds).

" $\geq$ ": Put $k = n - h$. We will show that $\sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} \max_i |log(\lambda_i(T(\mathbb{M}))) - log(\lambda_i(T(\mathbb{M}_k)))| < \infty$, or that only $n - h$ arbitrary points can not carry the estimate beyond all bounds.

First, we prove that the estimate of scatter cannot be infinitely small, or that there exists $Q_2 > 0$ such that $\lambda_i(T(\mathbb{M}_k)) > Q_2$ for all $i \in \{1, \ldots, p\}$ and all $\mathbb{M}_k \in \mathbb{M}^{(k)}$. Let $\mathbb{M}_k \in \mathbb{M}^{(k)}$. We know that if we have $n - h$ arbitrary points, then in any h columns of $\mathbb{M}_k$ will be at least $p + 1$ "good" columns, that are not arbitrary (thus lying in a general position). The rest of this claim can be done using Lemma 5. There exists Q such that every eigenvalue of the sample covariance matrix computed from those h points will be at least Q. If we take all $\binom{n}{p+1}$ choices of choosing $p + 1$ columns, for every choice there exists such $Q > 0$ and we choose the minimal one to be Q_2. Then, for every h columns of every $\mathbb{M}_k \in \mathbb{M}^{(k)}$, the smallest eigenvalue of a sample covariance matrix (computed from those h columns) will be greater than Q_2, what we wanted to show.

Finally, we prove that the estimate of scatter cannot be carried infinitely far, or that there exists $Q_1 \in \mathbb{R}$ such that $\lambda_i(T(\mathbb{M}_k)) < Q_1$ for all $i \in \{1, \ldots, p\}$ and all $\mathbb{M}_k \in \mathbb{M}^{(k)}$. For this, it is sufficient to show that there exists Q_1 such that $det(T(\mathbb{M}_k)) \leq Q_1$ (because determinant is only a multiple of the eigenvalues, and all eigenvalues are greater than Q_2). Let $\mathbb{M}_k \in \mathbb{M}^{(k)}$. Let us denote H the set of those h "good" points, that are the same as in matrix $\mathbb{M}$. Denote Q_H the determinant of the sample covariance matrix computed only from points from

H. From the definition of the MCD estimator, we take for our estimate those h points that have the minimum determinant of the covariance matrix, therefore $det(T(\mathbb{M}_k)) \leq Q_H$. Again, we want to show this for every initial choice of h columns. If we take all $\binom{n}{h}$ choices of choosing H, then for every H there exists such $Q_H < \infty$. If we take $Q_1 := \max Q_H$, then for every $\mathbb{M}_k$ holds $det(T(\mathbb{M}_k)) \leq Q_1$, what we wanted to show.

We concluded that $\forall \mathbb{M}_k \in \mathbb{M}^{(k)}$, $\forall i \in \{1, \dots, p\} : Q_1 > \lambda_i(T(\mathbb{M}_k)) > Q_2$, therefore it holds $\sup_{\mathbb{M}_k \in \mathbb{M}^{(k)}} \max_i |log(\lambda_i(T(\mathbb{M}_k))) - log(\lambda_i(T(\mathbb{M})))| < \infty$. □

3.2.4 Efficiency

It is well known that when $X_1, \dots, X_n \sim N_p(\mu, \Sigma)$, then the unbiased estimators of mean and variance with the minimal variance are $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ and $\mathbb{S}_n = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)(x_i - \overline{x}_n)^t$. We can estimate the variance of the MCD estimator using a simulation study. For this simulation, we took $n = 200$ random points from a multivariate normal distribution and computed their MCD estimates of location and scatter. We approximated their efficiency by dividing the sample variance of these estimates by the sample variance of the sample mean (or the sample variance matrix). This result is random, therefore in order to obtain more accurate results, we repeated this step 100 times. Means of these 100 results and their standard deviations are in Table 3.1. In this table, we show only the results for the scatter, efficiency for the location parameter was similar. The source code of this simulation (in programming language R) is in Attachment B.

Theorem 7. *The MCD estimator has better efficiency in greater dimensions. Moreover, it holds* $\lim_{p\to\infty} \frac{var\overline{X}}{var(\widehat{\mu}_{MCD})} = \frac{h}{n}$.

Proof. See [2]. □

It is obvious that by increasing h we obtain greater efficiency. As we can see in Table 3.1, the efficiency is insufficient for low p. Due to this property, in programming tools such as Matlab, or R has in the implementation of the MCD estimator the default value $h = 0.75n$, because it is a reasonable compromise between robustness and efficiency.

Efficiency			
h/n	p	Classical MCD estimator	Reweighted MCD estimator
0.5	2	6% (sd=4.4%)	48% (sd=3.1%)
0.5	10	18% (sd=4.3%)	81% (sd=2.1%)
0.75	2	26% (sd=3.9%)	49% (sd=3.4%)
0.75	10	45% (sd=3.1%)	82% (sd=1.7%)

Table 3.1: Efficiency of the MCD estimator and of the reweighted MCD estimator for selected values of dimensions p and quotients h/n. Values are obtained numerically from computing 100 repetitions (R script in Attachment B), therefore are not exact (sd stands for the standard deviation of the sampled results).

3.3 Reweighted MCD estimator

In order to obtain a better efficiency of an estimator, we can use the reweighted MCD estimator of location and scatter defined as follows:

$$\begin{aligned}
\widehat{\mu}_{RMCD} &= \frac{\sum_{i=1}^{n} W(d_i^2) x_i}{\sum_{i=1}^{n} W(d_i^2)}, \\
\widehat{\Sigma}_{RMCD} &= \frac{c_1}{n} \sum_{i=1}^{n} W(d_i^2)(x_i - \widehat{\mu}_{RMCD})(x_i - \widehat{\mu}_{RMCD})^t,
\end{aligned}$$

where $d_i = \sqrt{(x_i - \widehat{\mu}_{MCD})^t \widehat{\Sigma}_{MCD}^{-1} (x_i - \widehat{\mu}_{MCD})}$, c_1 is an appropriate consistency factor and $W(x)$ is an appropriate weight function. In `R` or `Matlab` is the default choice for the weight function $W(x) = 1(d_i^2 \leq \chi^2_{p,0.99})$, that can be considered as a function providing a cut-off for the outlying observations found by the classical MCD estimate.

The reweighted MCD estimator is affine equivariant, robust with high breakdown value and with better efficiency than the classical MCD estimator. Proofs of these results are analogous to the proofs for the classical MCD estimator [13]. Therefore, the reweighted MCD estimator is even more commonly used in practice, and in `R` it is used as the default MCD estimator.

4. MVE estimator

There exist other methods for finding a robust estimate of location and scatter parameters [14]. For example, there are M-estimators, which are a generalization of the maximum likelihood estimators. They are usually defined for a more general class of distributions, they have properties different from those of the MCD estimator, and usually does not handle such a great number of outliers. We will introduce one other method called MVE.

The minimal volume ellipsoid (MVE) estimator is similar to the MCD estimator in many ways. But in this case, we will find an ellipsoid with a minimal volume containing at least h observations and compute location and scatter only from those observations.

As we proved in Theorem 1 and Theorem 2, we know that $\{x \in \mathbb{R}^p : (x-t)^t \mathbb{C}^{-1}(x-t) \leq c_1\}$ forms an ellipsoid for $\mathbb{C} \in PDS(p)$. Moreover, we know the shape of this ellipsoid. The volume of such an ellipsoid is $V(E) = K_p det(\mathbb{C})$ for some constant $K_p \in \mathbb{R}$ for every dimension $p \in \mathbb{N}$.

We start with a lemma which will guarantee the existence and uniqueness of the MVE estimator.

Lemma 6. *Let $c_1 \in \mathbb{R}$, $n, p \in \mathbb{N}$, such that $c_1 > 0$, $n > p$, and let $x_1, \ldots, x_n \in \mathbb{R}^p$ lie in a general position. Then there exists a unique ellipsoid $E = \{x \in \mathbb{R}^p : (x-t)^t \mathbb{C}^{-1}(x-t) \leq c_1\}$ which contains $x_1, \ldots, x_n$ and such that $det(\mathbb{C}) < det(\mathbb{C}_2)$ for every other ellipsoid $E_2 = \{x \in \mathbb{R}^p : (x-t_2)^t \mathbb{C}_2^{-1}(x-t_2) \leq c_1\}$ which contains $x_1, \ldots, x_n$. Moreover, at least $p+1$ points from $x_1, \ldots, x_n$ lie on the surface of E. Therefore, E is the minimum volume ellipsoid for the points on its surface (ellipsoid uniquely determined by those $p+1$ points).*

Proof. The proof of this theorem requires deeper knowledge of measure theory. It may be found in [15] or in [16]. □

Definition. *Let $c_1 > 0$. The raw MVE location estimator $\widehat{\mu}_{MVE}$ and scatter estimator $\widehat{\Sigma}_{MVE}$ with parameter $h \in \mathbb{N}$, where $n/2 + 1 \leq h \leq n$, minimize the determinant of $\mathbb{C}$ subject to the condition*

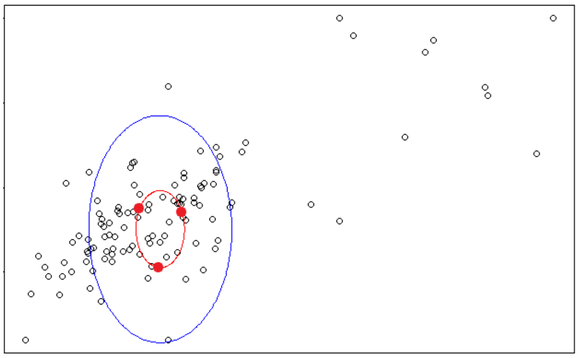
$$\#\{i : (x_i - t)^t \mathbb{C}^{-1}(x_i - t) \leq c_1\} \geq h,$$

where the minimization is over all $t \in \mathbb{R}^p$ and $\mathbb{C} \in PDS(p)$.

In other words, we choose as our estimate of the scatter matrix $\widehat{\Sigma}_{MVE}$ the matrix with the minimal determinant for which there exists $t \in \mathbb{R}^p$ such that the ellipsoid $\{x \in \mathbb{R}^p : (x-t)^t \widehat{\Sigma}_{MVE}^{-1}(x-t) \leq c_1\}$ contains at least h observations.

The consistency factor c_1 is usually chosen so that $\widehat{\Sigma}_{MVE}$ is a consistent estimator of the covariance matrix for normal data, in which case we put $c_1 = \chi^2_{p,q}$, where $q = \lim_{n \to \infty} h/n$. This result can be achieved by evaluating the condition $\#\{i \leq n : (x_i - \mu)^t \Sigma^{-1}(x_i - \mu) \leq c_1\} \geq h$ for true values μ, Σ instead of the estimated ones. Using Lemma 1 we can conclude that the condition holds if and only if $c_1 = \chi^2_{p,q}$ (for $n \to \infty$).

Figure 4.1: Computation of the MVE algorithm. The red $p+1$ points represent the first randomly chosen set. The red ellipse is the corresponding minimum ellipse. The blue ellipse represents the inflated red ellipse such that it contains $h = 0.75n$ observations.



4.1 MVE algorithm

Computing all $\binom{n}{h}$ evaluations in the MVE estimator is computationally expensive, so we rather satisfy with an approximate solution. We limit our search by taking not h subsets, but only $(p+1)$ subsets, which is computationally more convenient. Then we deflate or inflate the found ellipsoid corresponding to these elements (i.e. multiply $\mathbb{C}$ by a constant) until it contains h elements.

This process can be seen in Figure 4.1. The algorithm for one randomly generated dataset is the following:

1. Take random $(p+1)$ elements (if their sample covariance matrix is singular, we add elements until it is non-singular) and compute their sample mean $\overline{x}_{p+1}$ and sample covariance matrix $\mathbb{S}_{p+1}$;

2. Compute for each $i \leq n$ the quantity $D_i := MD_{\overline{x}_{p+1},\mathbb{S}_{p+1}}(x_i)$ and denote by D the h-th smallest squared distance of all D_i;

3. Denote $f = D/c_1$ (where $c_1 = \chi^2_{p,q}$ with $q = h/n$) and $\beta = f^{p/2}det(\mathbb{S}_{p+1})^{1/2}$;

4. Return $\overline{x}_{p+1}$ and $\beta\mathbb{S}_{p+1}$.

Note that $f^{p/2}det(\mathbb{S}_{p+1})^{1/2} = [det(f\mathbb{S}_{p+1})^{1/2}]$, which stands for the volume of the ellipsoid corresponding to $\overline{x}_{p+1}$ and $\mathbb{S}_{p+1}$ multiplied by a scaling factor. It is important to notice that this algorithm does not return the optimal solution, only when an appropriate first $(p+1)$ subset is chosen. Anyway, repeating this algorithm for each $\binom{n}{p+1}$ evaluations is still computationally expensive, so only a random collection is chosen, usually with 3000 or 30000 random initial $(p+1)$-subsets (there are surely better options for the number of initial subsets, this is a compromise for large and for small number of observations n).

4.2 MVE properties

We will show that the MVE estimator is an affine equivariant estimator with the same breakdown value as the MCD estimator. On the other hand, what makes it a less attractive choice is its efficiency.

Theorem 8. *The MVE estimator of location and scatter with parameter h is affine equivariant with the breakdown point $\frac{1}{n}min\{n-h+1, h-p\}$.*

Proof. Affine equivariance follows from the fact that if a point is in the ellipsoid, then after any affine transformation stays in the ellipsoid. Likewise, if a point is not in the ellipsoid after any affine transformation stays outside of the ellipsoid. Therefore, the ellipsoid with a minimal volume containing at least h points stays being the ellipsoid with a minimal volume containing at least h points after any affine transformation.

The proof of the breakdown point is analogous to that for the MCD estimator in Section 3.2.3. Both are discussed in [8]. □

Theorem 9. *Let $X_1, X_2, \ldots$ be independent random vectors, $X_i \sim N_p(\mu, \Sigma)$. Then $n^{\frac{1}{3}}(\widehat{\Sigma}_{MVE} - \Sigma) \xrightarrow[n\to\infty]{D} T$, where T is a non-degenerative random vector described in [16]. Therefore, the asymptotic efficiency of the MVE estimator is* $0\,\%$*.*

Proof. See [16]. □

There exist several improvements of the MVE that increase its efficiency. See [17], but usually if they improve the efficiency, then some other property is compromised, such as the breakdown point or the time complexity of the algorithm.

Remark (Comparison of MCD and MVE)**.** *As we have shown, the MVE estimator has lower efficiency than the MCD estimator. The reason is that it loses some information during the process — consider two data sets in Figure 4.2. The MVE scatter matrix will be similar for both, but the MCD scatter matrix of the first data set will have a greater determinant than for the second one (for h large enough so that the whole ellipse of points on the left panel will be considered in the "best" h points). The main difference is, that for the MVE is not important how are the observations in the ellipsoid distributed.*

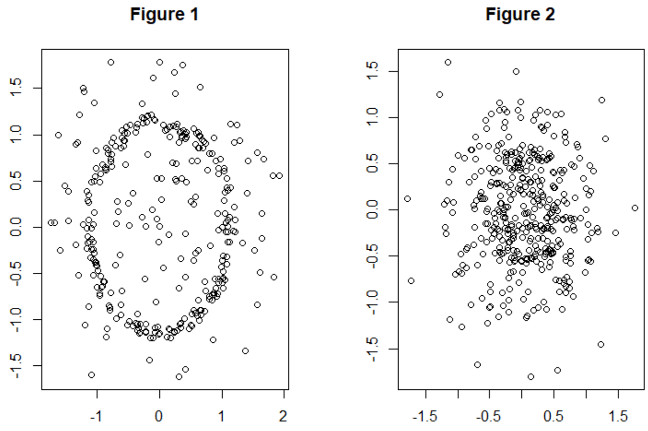


Figure 4.2: The fundamental difference between MCD and MVE — MVE does not distinguish between the two different distributions inside the tolerance ellipsoid, but MCD estimator does.

Conclusion

In this thesis, we introduced two approaches for finding robust estimators of location and scatter parameters in elliptically symmetric distributions.

At first, we described a method for finding a tolerance region using Mahalanobis distances. We showed, that this region forms an ellipsoid, and described its shape and derived its properties.

Second, we introduced the MCD estimator of location and scatter. This estimator uses only a fraction of observations with the minimal determinant of their sample covariance matrix. We showed the difference between using this robust estimate and using classical methods. Moreover, we used the MCD estimator in an application.

After that, we discussed the properties of the MCD estimator. First, we introduced the affine equivariance and the breakdown point. Then we proved that the MCD estimator is affine equivariant, with the highest possible breakdown point that can be achieved for an affine equivariant estimator. It can handle up to $\frac{\lfloor\frac{n-p+1}{2}\rfloor}{n}$ of outlying observations. We discussed the consistency and efficiency of the MCD estimator, where we used a simulation study rather than giving a proper proof. Moreover, we introduced the reweighed improvement of the MCD estimator and briefly discussed the algorithm for its computation.

Finally, we introduced the MVE estimator of location and scatter. This estimator fits the ellipsoid with the smallest volume containing a given fraction of the observations. We discussed the algorithm for its computation and listed some of its properties.

In the end, we can summarize that both MCD and MVE estimators are affine equivariant, robust with the highest possible breakdown point, and they are both consistent estimators of location and scatter. The MVE estimator has worse efficiency, but it is maybe a more natural choice as it can be more easily conceived. We conclude that these estimators are both very effective alternatives to the classical methods.

A. Attachment: Basic theorems from linear algebra and the table of notations

Proofs of theorems A1–A4 can be found in [18].

Theorem A1: [SVD decomposition] Suppose we have a real matrix $\mathbb{M} \in \mathbb{R}^{m \times n}$. Then, there exists a factorization, called the *singular value decomposition* of $\mathbb{M}$, of the form $\mathbb{M} = \mathbb{U}\mathbb{D}\mathbb{V}^t$, where $\mathbb{U}$ is an $m \times m$ orthogonal matrix, $\mathbb{D}$ is a diagonal $m \times n$ matrix (elements that are not on the main diagonal are 0) with non-negative real numbers $\mathbb{D}_{i,i}$ on the diagonal, $\mathbb{V}$ is an $n \times n$ orthogonal matrix. In case $\mathbb{M} \in PDS(n)$, then $\mathbb{U} = \mathbb{V}$ and $\forall i \in \{1, \ldots, n\} : \mathbb{D}_{i,i} > 0$. Moreover, $\mathbb{D}_{i,i}$ are the eigenvalues of matrix $\mathbb{M}$.

Theorem A2: [about orthogonal matrix] Let $x, y \in \mathbb{R}^p$ and $\mathbb{U} \in \mathbb{R}^{p \times p}$ be an orthogonal matrix. Then $||x|| = ||\mathbb{U}(x)||$ and $\sphericalangle[x, y] = \sphericalangle[\mathbb{U}(x), \mathbb{U}(y)]$, where $\sphericalangle[a, b]$ denotes the angle between vectors a, b. Specially, for an orthonormal basis $u_1, \ldots, u_p$ is also $\mathbb{U}(u_1), \ldots, \mathbb{U}(u_p)$ an orthonormal basis.

Theorem A3: [square root of a matrix] For any matrix $\mathbb{M} \in PDS(p)$ there exists a non-singular matrix $\mathbb{A} \in \mathbb{R}^{p \times p}$, such that $\mathbb{M} = \mathbb{A}\mathbb{A}^t$.

Theorem A4: [matrix norm] Let $x \in \mathbb{R}^p$, $\mathbb{A}, \mathbb{M} \in \mathbb{R}^{p \times p}$. Let $||\mathbb{A}|| := \max\{||\mathbb{A}y|| : y \in \mathbb{R}^p, ||y|| = 1\}$ denote a matrix norm. Then it holds $||\mathbb{A}x|| \leq ||\mathbb{A}||\,||x||$ and $||\mathbb{A}\mathbb{M}|| \leq ||\mathbb{A}||\,||\mathbb{M}||$. Moreover, if $\mathbb{A} \in PDS(p)$, and we denote λ_i the i-th greatest eigenvalue of $\mathbb{A}$, then it holds $\lambda_1||x|| \geq ||\mathbb{A}x|| \geq \lambda_p||x||$.

Theorem A5: [sample mean and sample covariance matrix] Let $n > p$ and let $x_1, \ldots, x_n \in \mathbb{R}^p$ be in general position (no more than p points lie on any hyper plane of dimension less than p). Denote $\overline{x}_n$ the sample mean, $\mathbb{S} = \frac{\sum_{k=1}^n (x_k - \overline{x}_n)(x_k - \overline{x}_n)^t}{n-1}$ the sample covariance matrix. Then it holds that $\mathbb{S}, \mathbb{S}^{-1} \in PDS(p)$.

Let $y \in \mathbb{R}^p$. The function $f(c) = \sum_{k=1}^n ((x_k - c)^t y)^2$ has a global minimum in $\overline{x}_n$. (With a special case $f(c) = \sum_{k=1}^n (x_k - c)^2$).

Proof. Proof that $\mathbb{S} \in PDS(p)$ is a special case of the proven Lemma 4 in Section 3.1.1. An inverse of a PDS matrix is also a PDS matrix [18]. The statement about the minimum follows from the fact that f is convex and $\nabla f(\overline{x}_n) = 0$. This is true, because $\frac{\partial f(x)}{\partial x_i} = \sum_{k=1}^n (x_k^{(i)} - c^{(i)}) 2y^{(i)} = 2y^{(i)} n (\frac{\sum_{k=1}^n x_k^{(i)}}{n} - c^{(i)})$ is zero for $c = \overline{x}_n$. Here, $\frac{\partial f(x)}{\partial x_i}$ denotes the partial derivative of f with respect to i, and $x_j^{(i)}$ denotes the i-th element of the vector x_j. □

Table A.1: Table of notations and basic definitions.

X	$\triangleq$	d-variate random vector for $d \geq 1$
F_X	$\triangleq$	distribution function of X
EX or μ	$\triangleq$	expected value of X
$Var(X)$ or Σ	$\triangleq$	variance of X
$N(\mu, \sigma^2)$	$\triangleq$	normal distribution with expected value μ and variance σ^2
$N_p(\mu, \Sigma)$	$\triangleq$	p-dimensional normal distribution with expected value μ and variance matrix Σ
$X \sim N(0,1)$	$\triangleq$	X has standard normal distribution
$\Phi_p(x)$	$\triangleq$	distribution function of p-dimensional standard normal distribution
u_q	$\triangleq$	q-quantile of the standard normal distribution
χ^2_p	$\triangleq$	chi-squared distribution with p degrees of freedom
$\chi_{p,q}$ or $\chi_p(q)$	$\triangleq$	q-quantile of the chi-distribution with p degrees of freedom, equally root of the q-quantile of the chi-squared distribution with p degrees of freedom
$\gamma(k, x)$	$\triangleq$	lower incomplete gamma function, defined as $\gamma(k,x) = \int_0^x t^{k-1} e^{-t} dt$
$\Gamma(k)$	$\triangleq$	gamma function, defined as $\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt$
$\overline{X}_n$	$\triangleq$	$\frac{\sum_{k=1}^n X_k}{n}$, called the sample mean, often written without index n
$\mathbb{S}_n$	$\triangleq$	$\frac{\sum_{k=1}^n (X_k - \overline{X}_n)(X_k - \overline{X}_n)^t}{n-1}$, called the sample covariance matrix, often written without index n
$\mathbb{N}$	$\triangleq$	set of all natural numbers
$\mathbb{A}, \mathbb{B}, \ldots$	$\triangleq$	matrix notation
$\mathbb{A}^t$	$\triangleq$	transposed matrix
$\mathbb{A}_{i,j}$	$\triangleq$	the value on the i-th column and j-th row of matrix $\mathbb{A}$
$\mathbb{I}_n$	$\triangleq$	n-dimensional identity matrix
$\lvert\lvert\mathbb{A}\rvert\rvert$	$\triangleq$	matrix norm, $\lvert\lvert\mathbb{A}\rvert\rvert = \max\{\lvert\lvert\mathbb{A}x\rvert\rvert : x \in \mathbb{R}^p, \lvert\lvert x\rvert\rvert = 1\}$
$PDS(p)$	$\triangleq$	class of positive definite symmetric matrices of dimension p
PDS	$\triangleq$	class of positive definite symmetric matrices of any dimension
$\nabla f(x)$	$\triangleq$	gradient of the function f in the point x
$\lfloor . \rfloor$	$\triangleq$	floor function
$\lceil . \rceil$	$\triangleq$	ceiling function
$1_A(x)$	$\triangleq$	indicator function of the set $\mathbb{A}$, returns 1 for $x \in A$, otherwise returns 0
standard basis of $\mathbb{R}^p$	$\triangleq$	orthonormal basis $(1,0,\ldots,0)^t, \ldots, (0,0,\ldots,1)^t$.
$\#i : condition$	$\triangleq$	number of i fulfilling condition

B. Attachment: R source code

```
#Computing efficiency of the MCD estimator
#Using "mvtnorm" and "robustbase" package

p=2 #dimension
a=0.75 #a=lim h/n
reweighted=FALSE

#########################
pocetopakovani=200;n=200;
library(mvtnorm)
library(robustbase)
listhodnot=1; listhodnot2=1;


for (i in 1:pocetopakovani) {

  x=rmvnorm(n, rep(0,p), diag(p))
  sigma=covMcd(x, alpha = alpha, raw.only = !reweighted )
  listhodnot=rbind(listhodnot, sigma$cov[1,1])
  listhodnot2=rbind(listhodnot2, var(x)[1,1])
}

var(listhodnot2)/var(listhodnot);
```

Figure B.1: R script for computing the efficiency of the MCD estimator of scatter.

Bibliography

[1] M. Hubert, P. J. Rousseeuw, and K. van Branden. ROBPCA: A new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

[2] R. W. Butler, P. L. Davies, and M. Jhun. Asymptotics for the minimum covariance determinant estimator. *Ann. Statist.*, 21(3):1385–1400, 1993.

[3] A. J. Wilson. Volume of n-dimensional ellipsoid. *Sciencia Acta Xaveriana*, 1(1):101–106, 2014.

[4] C. Becker. *Robustness and Complex Data Structures. Festschrift in Honour of Ursula Gather.* Springer Berlin Heidelberg, 2013.

[5] http://cms.enviroportal.sk/odpady/verejne-informacie.php, accessed 16.3.2019.

[6] P. J. Rousseeuw and K. van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[7] K. Sommerová. Exploiting numerical linear algebra to accelerate the computation of the MCD estimator. Master thesis, Charles University, Faculty of Numerical analysis, 2018.

[8] H. P. Lopuhaä and P. J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.*, 19:229–248, 1991.

[9] J. Anděl. *Základy matematické statistiky.* Matfyzpress, Praha, 2005.

[10] S. András and A. Baricz. Properties of the probability density function of the non-central chi-squared distribution. *J. Math. Anal. Appl.*, 346(2):395–402, 2008.

[11] C. Croux and G. Haesbroeck. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *J Multivariate Anal*, 71(2):161–190, 1999.

[12] R. W. Butler. Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule. *Ann. Statist.*, 10(1):197–204, 1982.

[13] H. P. Lopuhaä. Asymptotics of reweighted estimators of multivariate location and scatter. *Ann. Statist.*, 27(5):1638–1665, 1999.

[14] Omelka M. Modern statistical methods. (lecture notes, MFF UK 2019), available online, accessed at 1.5.2019 http://www.karlin.mff.cuni.cz/omelka/Soubory/nmst434/nmst434_course-notes.pdf.

[15] B. W. Silverman and D. M. Titterington. Minimum covering ellipses. *SIAM J. Sci. Stat. Comput.*, 1(4):401–409, 1980.

[16] L. Davies. The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.*, 20(4):1828–1843, 1992.

[17] C. Croux and G. Haesbroeck. An easy way to increase the finite-sample efficiency of the resampled minimum volume ellipsoid estimator. *Comput. Statist. Data Anal.*, 25(2):125–141, 1997.

[18] L. Barto and J. Tůma. Lineární algebra, (lecture notes, MFF UK 2019), available online, accessed at 1.5.2019. http://www.karlin.mff.cuni.cz/ barto/LinAlg/skripta_la6.pdf.