Probabilistic algorithms for computing the LTS estimate.

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Bachelor's thesis

Probabilistic algorithms for computing the LTS estimate

Martin Jenč

Department of Applied Mathematics
Supervisor: Ing. Karel Klouda, Ph.D.

April 29, 2019

Acknowledgements

THANKS to everybody

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as school work under the provisions of Article 60(1) of the Act.

In Prague on April 29, 2019 .....

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis

Jenč, Martin. *Probabilistic algorithms for computing the LTS estimate*. Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2019.

Abstrakt

Metoda nejmenších usekaných čverců je robustní verzí známé metdoy nejmenších čverců, jednou ze základních metod regresní analýzy, používané k odhadování koeficientů lineárního regresního modelu. Výpočet odhadadu pomocí metody nejmenších usekaných čtverců je znám jako NP-težký a proto jsou v praxi nejčatěji používány pouze subotimalní pravděpodobnostní algoritmy. Mimo popisu těchto algoritmů navhrneme několik způsobů jak je zkobinovat za účelem dosažení lepších výslekdů.

Klíčová slova nejmenší usekané čtverce, LTS odhad, linearní regrese, robustní statistika, nejmenší čverce, chybná pozorovaní

Abstract

The least trimmed squares method is a robust version of the method of least squares which is an essential tool of regression analysis used to find an estimate of coefficients in the linear regression model. Computing the least trimmed squared estimate is known to be NP-hard, and hence only suboptimal probabilistic algorithms are usually used in practice. Besides describing those algorithms, we propose a few ways of combining those algorithms to obtain better performance.

Keywords least trimmed squares, LTS estimate , linear regression, robust statistics, ordinary least squares, outliers

Contents

List of Figures

Introduction

Least trimmed squares (LTS) is one of many modifications very well known method ordinary least squares (OLS). Both of these methods are tools of regression analysis, which is a group of processes used to estimate the dependence of variables. Specifically in the case when we try to estimate dependence one variable on one or multiple others. The regression analysis uses the regression models, and in the case of OLS and LTS methods, that model is the linear regression model.

To OLS estimate produces reliable results, many strong assumptions about the data have to be fulfilled. It is assumed that data is generated specific way as well as that the data does not contain measurement errors called outliers. Those assumptions are in practice hardly fulfilled, because outliers in the data are very common. The OLS estimate is in such cases unreliable.

Problems of classical statistic methods try to solve robust statistic. Its methods are usually emulations of classical statistic methods but try to provide good performance even if data contains a large number of outliers. That means such methods does not rely so much on assumptions which are difficult to achieve in practice. Because the OLS method is one of the essential tools of linear regression analysis, multiple its alternatives have been designed to fulfill the assumptions of a robust estimator.

One of the methods is LTS. Its idea is simple, but unlike the OLS method, the exact formula for calculation is not known, so that only to date known exact algorithms for computing the LTS estimate are combinatorial. For this reason beside exact algorithms, several probabilistic algorithms have been proposed.

This work is divided into three chapters. In the first chapter, we introduce theory required to understand linear regression model, OLS and LTS. We mention properties of both methods and also describe filed of its usage.

In the second chapter, we cover algorithms for calculating the OLS estimate. It is necessary because most of the algorithm used to compute the LTS estimate relies on those algorithms. Next, we describe all currently used

algorithms for calculating LTS estimate. They consist of several probabilistic and also few exact ones. In this chapter, we also show that those algorithms can be easily combined to obtain higher speed and performance.

In the last chapter, we describe our experimental results. At first, we cover data generator which is used for our experiments and which can provide data affected by various types of outliers. Next, we provide information about our implementation of all algorithms for chapter two. Finally, we present our results for specific data.

Chapter 1

Least trimmed squares

In this chapter, we introduce one of the most common regression analysis models which is known as the linear regression model. It aims to model the relationship between one variable which is called *dependent* and one or more variables which are called *explanatory*. The relationship is based on a model function with parameters which are not known in advance and are to be estimated from data. We also describe one of the most common methods for finding those parameters in this model, namely the ordinary least squares method. It is important to note that it is usual to consider vectors as column vectors. On the other hand, we denote a row vector as a transposed vector.

1.1 Linear regression model

Definition 1. The *linear regression model* is

$$y = \boldsymbol{x}^T \boldsymbol{w} + \varepsilon \tag{1.1}$$

where $y \in \mathbb{R}$ is random variable called *dependent variable*, vector $\boldsymbol{x}^T = (x_1, x_2, \ldots, x_p)$ is column vector of *explanatory variables*. Usually we call x_i a regressor. Finally $\varepsilon \in \mathbb{R}$ is a random variable called *noise* or *error*. Vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_p)$ is vector of parameters called *regression coefficients*.

In practice, we usually deal with multiple dependent variables so we define multiple linear regression model.

Definition 2. The *multiple linear regression model* is

$$y_i = \boldsymbol{x_i}^T \boldsymbol{w} + \varepsilon_i, i = 1, \ldots, n. \tag{1.2}$$

It is common to write the whole model in a matrix form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon} \tag{1.3}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_1}^T \\ \boldsymbol{x_2}^T \\ \vdots \\ \boldsymbol{x_n}^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

That means that we can think of rows of matrix $\boldsymbol{X}$ as columns vectors x_i written into the row. This means that even when we think of all vectors as columns given the matrix $\boldsymbol{X}$ we consider x_i as row vectors.

Usually we assume that errors are independent and identically distributed so that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$.

It this work we talk about the multiple linear regression model, but for simplicity, we call it merely the linear regression model.

Note 3. It is usual refer to given $\boldsymbol{X}$ and y as to ***data set*** and to y_i with corresponding $\boldsymbol{x_i}$ as to ***data sample*** or ***observation***. Sometimes it's also useful to refer to multiple observations as to subset of observations.

1.1.1 Prediction with the linear regression model

The linear regression model contains the vector of actual regression coefficients which are unknown and which we need to estimate in order to be able to use the model for predictions. Let us assume that we already have estimated regression coefficients as $\hat{\boldsymbol{w}}$. Then the predicted values of y are given by

$$\hat{y} = \hat{\boldsymbol{w}}^T \boldsymbol{x}. \tag{1.4}$$

True value of y is given by

$$y = \boldsymbol{w}^{*T} \boldsymbol{x} + \varepsilon \tag{1.5}$$

where $\boldsymbol{w}^*$ represents actual regression coefficients which we estimate.

Because we assume linear dependence between dependent variable y and explanatory variables $\boldsymbol{x}$ then what makes y random variable is a random variable ε. Because we assume that $\mathbb{E}(\varepsilon) = 0$ we can see that

$$\mathbb{E}(y) = \mathbb{E}(\boldsymbol{x}^T \boldsymbol{w}) + \mathbb{E}(\varepsilon) = \mathbb{E}(\boldsymbol{x}^T \boldsymbol{w}) \tag{1.6}$$

so $\hat{y}$ is a point estimation of the expected value of y.

Intercept

In real world situations it is not usual that $\mathbb{E}(\varepsilon) = 0$. Consider this trivial example.

Example 4. Let us consider that y represents price of the room and $x \in \mathbb{N}$ represents the number of the windows in such a room. If this room does not have windows thus $x = 0$ and $\mathbb{E}(\varepsilon) = 0$ then $y = \boldsymbol{w}^T\boldsymbol{x} + \varepsilon$ equals zero. But it is very unlikely that room without windows is free.

Because of that, it is very common to include one constant regressor $x_0 = 1$ then the corresponding coefficient w_0 of $\boldsymbol{w}$ is called *intercept*. We refer this model as a *model with an intercept*. Intercept then corresponds to expected value of y when all regressors are zero and prevent the problem from example 4. Given that y is a random variable due to the ε, intercept actually corresponds to $\mathbb{E}(\varepsilon) = \mu$. With regards to this fact we can still assume that in the model with intercept $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$. In this work, we consider the model with the intercept unless otherwise specified. This means that we consider $\boldsymbol{x} = (x_1, x_2 \dots x_p)$ where constant $x_1 = 1$ represents intercept.

Note 5. Sometimes in model with intercept, the explanatory variable $\boldsymbol{x}$ is marked as $\boldsymbol{x} \in \mathbb{R}^{p+1}, \boldsymbol{x} = (x_0, x_1 \dots x_p)$ which means that actual observation $\boldsymbol{x} \in \mathbb{R}^p$ and the intercept $x_0 = 1$ is explicitly marked.

Some other assumptions about $\boldsymbol{\varepsilon}$ are in practice invalid, and we describe this in section 1.3.

1.2 Ordinary least squares

We want to estimate $\boldsymbol{w}$ so that the error of the model will the least possible. Measurement of this error done by a *loss function* $L : \mathbb{R}^2 \to \mathbb{R}$, which in case of ordinary least squares (OLS) is quadratic loss function $L(y, \hat{y}) := (y - \hat{y})^2$. So the basic idea is to find $\hat{\boldsymbol{w}}$ so that it minimizes the sum of squared *residuals*

$$r_i(\boldsymbol{w}) = y_i - \hat{y}_i = y_i - \boldsymbol{x_i}^T\boldsymbol{w}, \ i = 1, 2, \dots, n. \tag{1.7}$$

This is commonly know as residual sum of squares RSS

$$RSS(\boldsymbol{w}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{w}) = (y_i - \boldsymbol{w}^T\boldsymbol{x_i})^2. \tag{1.8}$$

Definition 6. The RSS as the function of $\boldsymbol{w}$ is an *objective function* for OLS denoted as

$$\mathrm{OF}^{(OLS,\boldsymbol{X},\boldsymbol{y})}(\boldsymbol{w}) = \sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x_i})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}) \tag{1.9}$$

The point of the minimum of this function

$$\hat{\boldsymbol{w}}^{(\boldsymbol{OLS},\boldsymbol{n})} = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min}\, \mathrm{OF}^{(OLS,\boldsymbol{X},\boldsymbol{y})}(\boldsymbol{w}) \tag{1.10}$$

is a definition of ordinary least squares estimate of true regression coefficients $\boldsymbol{w}^*$.

To find the minimum of this function, first, we need to find the gradient by calculating all partial derivatives

$$\frac{\partial \mathrm{OF}^{(OLS,\boldsymbol{X},\boldsymbol{y})}}{\partial w_j} = \sum_{i=1}^{n} 2(y_i - \boldsymbol{w}^T \boldsymbol{x_i})(-x_{ij}), \quad j \in \{1, 2, \ldots, p\}. \tag{1.11}$$

By this we obtain the gradient

$$\nabla \mathrm{OF}^{(OLS,\boldsymbol{X},\boldsymbol{y})} = -\sum_{i=1}^{n} 2(y_i - \boldsymbol{w}^T \boldsymbol{x_i})\boldsymbol{x_i}. \tag{1.12}$$

Putting gradient equal to zero we get the so called *normal equation*

$$-\sum_{i=1}^{n} 2(y_i - \boldsymbol{w}^T \boldsymbol{x_i})\boldsymbol{x_i} = 0 \tag{1.13}$$

that can be rewritten in matrix form as

$$\boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} = 0. \tag{1.14}$$

Let us now construct the hessian matrix using second-order partial derivatives:

$$\frac{\partial^2 \mathrm{OF}^{(OLS,\boldsymbol{X},\boldsymbol{y})}}{\partial w_h \partial w_j} = \sum_{i=1}^{n} 2(-x_{ik})(-x_{ij}), \quad h \in \{1, 2, \ldots, p\}. \tag{1.15}$$

We get

$$\boldsymbol{H}_{\mathbf{OF}^{(\boldsymbol{OLS},\boldsymbol{X},\boldsymbol{y})}} = 2\boldsymbol{X}^T \boldsymbol{X}. \tag{1.16}$$

We can see that hessian $\boldsymbol{H}_{\mathbf{OF}^{(\boldsymbol{OLS},\boldsymbol{X},\boldsymbol{y})}}$ is always positive semi-definite because for all $\boldsymbol{s} \in \mathbb{R}^p$

$$\boldsymbol{s}^T (2\boldsymbol{X}^T \boldsymbol{X})\boldsymbol{s} = 2(\boldsymbol{X}\boldsymbol{s})^T (\boldsymbol{X}\boldsymbol{s}) = 2 \left\| \boldsymbol{X}\boldsymbol{s} \right\| \tag{1.17}$$

It is easy to prove that twice differentiable function is convex if and only if the hessian of such function is positive semi-definite. Moreover, any local minimum of the convex function is also the global one. Give that the solution of (1.14) gives us the global minimum.

Assuming that $\boldsymbol{X}^T \boldsymbol{X}$ is a regular matrix, then its inverse exists, and solution can be explicitly written as

$$\hat{\boldsymbol{w}}^{(OLS,n)} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}, \tag{1.18}$$

where (OLS, n) denotes that $\hat{\boldsymbol{w}}$ is estimate of true regression coefficients $\boldsymbol{w}^*$ given by the OLS method for n observations.

Moreover, we can see that if $\boldsymbol{X}^T \boldsymbol{X}$ is a regular matrix, then the hessian $\boldsymbol{H}_{\mathbf{OF}^{(\boldsymbol{OLS},\boldsymbol{X},\boldsymbol{y})}}$ is positive definite so the $\mathrm{OF}^{(OLS,\boldsymbol{X},\boldsymbol{y})}$ is strictly convex thus $\hat{\boldsymbol{w}}^{(OLS,n)}$ is then strict global minimum.

1.2.1 Properties of OLS estimation

Gauss-Markov theorem tells us that under specific conditions (assumptions about the regression model) is OLS estimation unbiased and efficient. In other words, Gauss-Markov theorem states that OLS is the best linear unbiased estimator (BLUE). Efficient means that any other linear unbiased estimator has the same or higher variance than OLS. Those conditions are:

- Expected value of errors $\mathbb{E}(\varepsilon_i) = 0,\ i = 1, 2, \ldots, n.$
- Errors are independently distributed and uncorrelated thus $cov(\varepsilon_i, varepsilon_j),\ i, j = 1, 2 \ldots, n,\quad i \neq j$
- Regressors $\boldsymbol{x_i}, i = 1, 2, \ldots, n$ and corresponding errors $\varepsilon_i,\ i = 1, 2, \ldots, n$ are uncorrelated.
- All errors have same finite variance. This is known as *homoscedasticity*.

Note 7. As described in example 4 the first assumption is usually fulfilled if we use the model with intercept. Rest of the conditions, on the other hand, are rarely met.

There are also other theorems which describe properties of OLS under specific conditions, but they are out if the scope of this work.

1.3 Robust statistics

Standard statistic methods expect some assumptions to work properly and fail if those assumptions are not met. A robust statistic is statistics that produce acceptable results even when data are from some unconventional distributions or if data contains errors which are not normally distributed. We should be highly motivated to use such methods because in practice we are often forced to work with such data on which classical statistics methods provide poor results.

Such assumptions about OLS method are described in section 1.2.1. Before we explain what happens if those conditions are not met or met only partially let us describe one of the most common reasons why assumptions are false.

1.3.1 Outliers

We stated many assumptions that are required for ordinary least squares to give a good estimate of $\hat{\boldsymbol{w}}$. Unfortunately, in real conditions, these assumptions are often false so that ordinary least squares do not produce acceptable results. One of the most common reasons for false assumptions are abnormal observations called outliers.

Outliers are very common, and they are for instance erroneous measurements such as transmission errors or noise. Another common reason is that nowadays we are mostly given data which were automatically processed by computers. Sometimes we are also presented data which are heterogeneous in the sense that they contain data from multiple regression models. In some sense outliers are inevitable. One would say that we should be able to eliminate them by precise examination, repair or removal of such data. That is possible in some cases, but most of the data we are dealing with are too big to check, and even more often we do not know how such should look.

Moreover, in higher dimensions, it is complicated to find outliers. Some methods try to find outliers, but such methods are only partially efficient. Let us note that robust models are sometimes not only useful to create models that are not being unduly affected by the presence of outliers but also capable of identifying data which seems to be outliers.

We have some terminology to describe certain types of outliers. We use terminology from [1]. Let us have observation $(y_i, \boldsymbol{x_i})$. If observation is not outlying in any direction we call it *regular observation* . If it is outlying in $\boldsymbol{x_i}$ direction we call it *leverage point*. We have two types of leverage points. If $\boldsymbol{x_i}$ is outlying but $(y_i, \boldsymbol{x_i})$ follows liner pattern we call it *good leverage point*. If it does not follow such a pattern we call it *bad leverage point*. Finally if $(y_i, \boldsymbol{x_i})$ is outlying only in y_i direction, we call it a *vertical outliers*.

1.3.2 Measuring robustness

There are a couple of tools to measure the robustness of statistics. The most popular one is called *breakdown point*. Then there are *empirical influence function* and *influence function and sensitivity curve*. For the sake of simplicity, we describe only breakdown point right now.

Definition 8. Let T be a statistics, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be an n-dimensional random sample and $T_n(\boldsymbol{x})$ value of this statistics with parameter $\boldsymbol{x}$. The breakdown point of T at sample $\boldsymbol{x}$ can be defined using sample $\overline{\boldsymbol{x}}^{(k)}$ where we exchange k points from original sample $\boldsymbol{x}$ with random values x_i. We get $T_n(\overline{\boldsymbol{x}}^{(k)})$. Then the *breakdown point* is

$$\text{bdpoint}(T, \boldsymbol{x_n}) = \frac{1}{n} \min S_{T,\boldsymbol{x_n}}, \tag{1.19}$$

where

$$S_{T,\boldsymbol{x_n},D} = \left\{ k \in \{1, 2, \ldots, n\} : \sup_{\boldsymbol{x_{new}}^{(k)}} \left\| T_n(\boldsymbol{x}) - T_n(\boldsymbol{x_{new}}^{(k)}) \right\| = \infty \right\}. \tag{1.20}$$

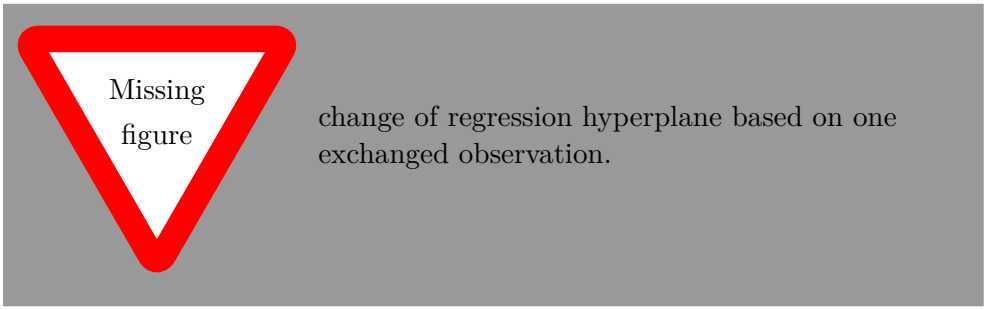
This definition is very general but let us specify it to our linear regression problem. It says that breakdown point is the function of the minimal number of observations needed to be changed for some others so then the estimator gives incorrect results. More robust estimators have higher breakdown point.

It is intuitive that reasonable breakdown point should not be higher than 0.5 [2] because if more than 50% of the data is exchanged, then the model of exchanged data overrides the model of the original data.

In the context of OLS estimator one outlier is enough to increase the value of T to any desired value [3] thus

$$\text{bdpoint}(OLS, \boldsymbol{x_n}) = \frac{1}{n}. \tag{1.21}$$

Figure [] gives us an example of how one outlier may change the hyperplane given by the OLS estimation of regression coefficients.



change of regression hyperplane based on one exchanged observation.

For an increasing number of data samples n this tends to zero. We can see that ordinary least squares estimator is not resistant to outliers at all. Due to this fact, multiple estimators similar to OLS have been proposed.

1.4 Least trimmed squares

The least trimmed squares (LTS) estimator is a more robust version of the OLS. In this section, we define LTS estimator and show that its breakdown point is variable and can go up to the maximum possible value of breakdown point, thus 0.5.

Definition 9. Let us have $\boldsymbol{X} \in \mathbb{R}^{n,p}$, $\boldsymbol{y} \in \mathbb{R}^{n,1}$, $\boldsymbol{w} \in \mathbb{R}^p$ and h, $n/2 \leq h \leq n$. The objective function of LTS is then denoted as

$$\text{OF}^{(LTS,h,n)}(\boldsymbol{w}) = \sum_{i=1}^{h} r^2_{i:n}(\boldsymbol{w}) \tag{1.22}$$

where $r^2_{i:n}$ denotes ith smallest squared residuum out of n thus $r^2_{1:n} \leq r^2_{2:n} \ldots r^2_{n:n}$ Even though that objective function of LTS seems similar to the OLS objective function, finding minimum is far more complex because the order of the least squared residuals is dependent on $\boldsymbol{w}$. Moreover $r^2_{i:n}(\boldsymbol{w})$ residuum is not uniquely determined if more squared residuals have same value as $r^2_{i:n}(\boldsymbol{w})$. This makes from finding LTS estimate non-convex optimization problem and finding global minimum is NP-hard [4].

1.4.1 Discrete objective function

The LTS objective function from definition 9 is not differentiable and not convex, so we are unable to use the same approach as with OLS objective function. Let us transform this objective function to a discrete version which can appropriately be used by algorithms to minimize it.

Let us assume for now that we know the $\hat{\boldsymbol{w}}^{(\boldsymbol{LTS},\boldsymbol{h},\boldsymbol{n})}$ vector of estimated regression coefficients. (LTS, h, n) denotes that coefficients are estimated using LTS on the h subset of n data samples. With this in mind, let π be the permutation of $\hat{n} = \{1, 2, \ldots, n\}$ such that

$$r_{i:n} = r_{\pi(j)}, \quad j \in \hat{n}. \tag{1.23}$$

Put

$$Q^{(n,h)} = \{\boldsymbol{m} \in \mathbb{R}^n, m_i \in \{0,1\}, i \in \hat{n}, \sum_{i=1}^{n} m_i = h\}, \tag{1.24}$$

which is simply the set of all vectors $\boldsymbol{m} \in \mathbb{R}^n$ which contain h ones and $n - h$ zeros. Let $\boldsymbol{m}_{LTS} \in Q^{(n,h)}$ such that $m_j^{(LTS)} = 1$ when $\pi(j) \leq h$ and $m_j^{(LTS)} = 0$ otherwise. Then

$$\hat{\boldsymbol{w}}^{(\boldsymbol{LTS},\boldsymbol{h},\boldsymbol{n})} = \underset{\sum_{i=1}^{h} r_{i:n}^2(\boldsymbol{w})}{\arg\min} = \underset{\sum_{i=1}^{n} m_i^{(LTS)} r_i^2(\boldsymbol{w})}{\arg\min}. \tag{1.25}$$

This means that if we know the vector $\boldsymbol{m}_{LTS}$ than we can compute the LTS estimate as the OLS estimate with $\boldsymbol{X}$ and $\boldsymbol{Y}$ multiplied by the diagonal matrix $\boldsymbol{M}_{LTS} = diag(\boldsymbol{m}_{LTS}^T)$:

$$\hat{\boldsymbol{w}}^{(\boldsymbol{LTS},\boldsymbol{h},\boldsymbol{n})} = (\boldsymbol{X}^T \boldsymbol{M}_{LTS} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{M}_{LTS} \boldsymbol{y}. \tag{1.26}$$

In other words, finding the minimum of the LTS objective function can be done by finding the OLS estimate (1.26) for all vectors $\boldsymbol{m} \in Q^{(n,h)}$. Thus if we denote $\boldsymbol{X}_{\boldsymbol{M}} = \boldsymbol{M}\boldsymbol{X}$ and $\boldsymbol{y}_{\boldsymbol{M}} = \boldsymbol{M}\boldsymbol{y}$ which corresponds to h subsets of $\boldsymbol{X}$ and $\boldsymbol{Y}$, then as described in [5],

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \mathrm{OF}^{(LTS,h,n)}(\boldsymbol{w}) = \min_{\boldsymbol{w} \in \mathbb{R}^p} \sum_{i=1}^{h} r_{i:n}^2(\boldsymbol{w}) \tag{1.27}$$

$$= \min_{\boldsymbol{w} \in \mathbb{R}^p, \boldsymbol{m} \in Q^{(n,h)}} \sum_{i=1}^{n} m_i r_i^2(\boldsymbol{w}) \tag{1.28}$$

$$= \min_{\boldsymbol{m} \in Q^{(n,h)}} \Big(\min_{\boldsymbol{w} \in \mathbb{R}^p} \mathrm{OF}^{(OLS,\boldsymbol{M}\boldsymbol{X},\boldsymbol{M}\boldsymbol{y})}(\boldsymbol{w}) \Big) \tag{1.29}$$

$$= \min_{\boldsymbol{m} \in Q^{(n,h)}} \Big(\min_{\boldsymbol{w} \in \mathbb{R}^p} \|\boldsymbol{M}\boldsymbol{y} - \boldsymbol{M}\boldsymbol{X}\boldsymbol{w}\|^2 \Big). \tag{1.30}$$

Substituting $\boldsymbol{w}$ with (1.25) we get

$$\begin{aligned}\hat{\boldsymbol{w}}^{(\boldsymbol{LTS},\boldsymbol{h},\boldsymbol{n})} &= \min_{\boldsymbol{m}\in Q^{(n,h)}} \left(\left\| \boldsymbol{M}\boldsymbol{y} - \boldsymbol{M}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{y} \right\|^2 \right) \\ &= \min_{\boldsymbol{m}\in Q^{(n,h)}} \left(\left\| \boldsymbol{M}\boldsymbol{y} - \boldsymbol{M}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{y} \right\|^2 \right)\end{aligned}$$

We can easily see, that have the objective function with argument $\boldsymbol{m} \in Q^{(n,h)}$. This objective function we mark as

$$\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}) = \left\| \boldsymbol{M}\boldsymbol{y} - \boldsymbol{M}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{y} \right\|^2. \tag{1.31}$$

We can also clearly see that the minimizing this OF could by done straightforwardly by iterating over the $Q^{(n,h)}$ set. That means we transformed the objective function to the discrete space of $Q^{(n,h)}$.

Unfortunately, this set is vast, namely $\binom{n}{h}$, so this approach is infeasible on more massive data sets. Multiple algorithms were proposed to overcome this problem. Majority of them are probabilistic algorithms, but besides those, some exact algorithms were proposed.

Finally, let us point out some fact about the number h of non-trimmed residuals and how it makes least trimmed squares robust. The LTS reaches maximum breakdown point 0.5 at $h = [(n/2] + [(p+1)/2]$ [3](where [.] denotes largest integer function). That means that up to 50% of the data can be outliers. In practice, the portion of outliers is usually lower; if an upper bound on the percentage of outliers is known, h should be set to match this percentage.

Chapter 2

Algorithms

In the previous chapter, we cover the theoretical background required to implement algorithms that are in this chapter. We introduced the discrete version of the LTS objective function which minimum is equivalent to the continuous one. Finding the minimum of this function requires to find the correct h subset and then calculate the estimate of regression coefficients $\boldsymbol{w}$. To achieve this, we need to examine all h subsets. The exhaustive approach fails due to the exponential size of $Q^{(n,h)}$. So what are the possibilities?

2.0.1 First attempts

Initial algorithms were based on iterative removal data samples whose residuum had the highest value based on OLS fit on the whole dataset. Such attempts are known to be flawed [6] because the initial OLS fit can be already profoundly affected by outliers and we may remove data samples which represent the actual model.

Other algorithms were based purely on a random approach. On such algorithm is Random solution algorithm [7] which randomly selects l of h subsets and subsequently compute OLS fit on each of them and chooses fit with a minimum value of the objective function. Such approach is straightforward, but the probability of selecting at least one subset from l subsets which don't contain outliers thus has a chance of producing good result tends to zero for an increasing number of data samples n as we'll describe in detail in section 2.2.2.

Another very similar algorithm called Resampling algorithm introduced in [8]. It selects vectors from $Q^{(n,p+1)}$ instead of $Q^{(n,h)}$. This minor tweak has a higher chance to succeed. Mainly because the probability of selecting l subsets of size $p+1$ thus it is independent of n gives the nonzero probability of selecting at least one subset such that it does not include outliers (see section 2.2.2 for more details). Besides that the number of vectors in this set is significantly lower than in $Q^{(n,h)}$ (at least if h is conservatively chosen so that $h = [(n/2] + [(p+1)/2])$.

2.0.2 Strong and weak necessary conditions

Generating all possible h subsets is computationally exhaustive and relying on randomly generating "good" h-element subsets does not lead to reliable results. So what are our options? In [9] two criteria called *weak necessary conditions* and *strong necessary conditions* are introduced. They state necessary properties which some h subset must satisfy to be subset which leads to the global optimum of the LTS objective function. Let us introduce those two necessary conditions. For that, it is convenient not only to label a h subset of *non-trimmed* observations but also the complementary subset of not used observations. We'll refer to this complementary subset as to *trimmed* subset.

Definition 10. A h-element subset corresponding to $\boldsymbol{m} \in Q^{(n,h)}$ satisfies the *strong necessary condition* if for any vector $\boldsymbol{m_{swap}}$ which differs from $\boldsymbol{m}$ only by swapping an 1 with one 0, we get $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}) \leq \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m_{swap}})$. Thus the value of discrete LTS objective function cannot be reduced by swapping one non-trimmed observation with one trimmed observation.

Based on this fact an algorithm can be created. We'll discuss it in detail in section 2.3.

The second necessary condition is named *weak necessary condition*

Definition 11. A h-element subset corresponding to $\boldsymbol{m} \in Q^{(n,h)}$ satisfies the *strong necessary condition* if $r_i^2(\hat{\boldsymbol{w}}^{(\boldsymbol{OLS},\boldsymbol{MX},\boldsymbol{My})})$ for all trimmed observation is greater of equal to the greatest non-trimmed squared residuum $r_j^2(\hat{\boldsymbol{w}}^{(\boldsymbol{OLS},\boldsymbol{MX},\boldsymbol{My})})$.

Again, based on this criteria an algorithm can be created. The corollary of this criteria together with proof can be found in section 2.2.

The interesting consequence which we'll use later gives us the following lemma.

Lemma 12. The strong necessary condition is not satisfied unless the weak necessary condition is satisfied. Thus if a strong condition is satisfied then weak is also.

Proof. We make proof by contradiction. Let us assume that we have $\hat{\boldsymbol{w}}^{(\boldsymbol{OLS},\boldsymbol{MX},\boldsymbol{My})}$ where $\boldsymbol{m} \in Q^{(n,h)}$ for which strong necessary condition is satisfied, but the weak necessary condition is not. That means there exists $\boldsymbol{x_i}$ with y_i from non-trimmed subset and $\boldsymbol{x_j}$ and y_j from trimmed subset such that

$$r_j^2(\hat{\boldsymbol{w}}^{(\boldsymbol{OLS},\boldsymbol{MX},\boldsymbol{My})}) < r_i^2(\hat{\boldsymbol{w}}^{(\boldsymbol{OLS},\boldsymbol{MX},\boldsymbol{My})}).$$

Thus

$$\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}) > \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}) + r_j^2 - r_i^2 \tag{2.1}$$

Now we need to show that $\boldsymbol{m_{swap}}$ vector that is created by swapping jth observation from trimmed subset with ith observation from non-trimmed subset leads to

$$\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}) + r_j^2 - r_i^2 \geq \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m_{swap}}). \tag{2.2}$$

That is true because the value of $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m_{swap}})$ is minimum on the given subset of observations. That is, of course, contradiction with our assumption which says that strong necessary condition is already satisfied. □

Now we have covered all the necessary theoretical background, and it is time to introduce currently known algorithms for computing the LTS estimate.

2.1 Computing OLS

In this section, we describe a few of many methods that can be used to obtain $\hat{\boldsymbol{w}}^{(OLS,n)}$. Those methods are parts of the algorithms used to calculate the LTS estimate.

In the following algorithms, we describe its time complexity. Because matrix multiplication is the fundamental part of all the following algorithms, it is, therefore, appropriate to mention a few facts about the time complexity of matrix multiplication.

There are multiple algorithms for matrix multiplication, for example:

- Naive algorithm; its time complexity is $\mathcal{O}(n^3)$.
- Strassen algorithm; its time complexity is $\mathcal{O}(n^{2.8074})$ [10].
- Coppersmith-Winograd; its time complexity is $\mathcal{O}(n^{2.375477})$[11]

In practice, however, the naive algorithm is usually used. Even though some of those algorithms have been efficiently implemented and are known to be numerically stable (primarily variations of Strassen algorithm), their error bound is weaker than in case of the naive algorithm. For this reason, they are not used even when they might improve performance [12].

Moreover current implementations of the linear algebra namely LAPACK and more importantly BLAS level 3 routines implementing matrix-matrix multiplication which are widely used in such linear algebra software implement naive matrix-matrix multiplication [13].

For that reason we assume in this work that matrix multiplication is $\mathcal{O}(n^3)$. For not square matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}, \boldsymbol{B} \in \mathbb{R}^{n \times p}$ it is then$\mathcal{O}(mnp)$.

2.1.1 Computation using a matrix inversion

Solving OLS using matrix inversion can be done as follows:

1. Compute $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{y}$.
2. Find inversion of $\boldsymbol{A}$.
3. Multiply $\boldsymbol{A}^{-1}\boldsymbol{B}$.

Observation 13. Time complexity of computing the OLS estimate on $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$ using matrix inversion is $O(p^2 n)$.

Proof. First, we compute $\boldsymbol{A} = \boldsymbol{X^T X}$ and $\boldsymbol{B} = \boldsymbol{X^T y}$. That gives us $\mathcal{O}(p^2 n + pn)$. Next, we compute inversion $\boldsymbol{C} = \boldsymbol{A}^{-1}$ which gives us $\mathcal{O}(p^3)$. Finally, we compute $\hat{\boldsymbol{w}}^{(OLS,n)} = \boldsymbol{CB}$ which is $\mathcal{O}(p^2)$. Together we get $\mathcal{O}(p^2 n + pn + p^3 + p^2)$. $p^2 n$ and p^3 asymptotically dominates over the rest so

$$\mathcal{O}(p^2 n + pn + p^3 + p^2) \sim \mathcal{O}(p^2 n + p^3). \tag{2.3}$$

Moreover if we assume that $n \geq p$ and usually even $n \gg p$ we get $\mathcal{O}(p^2 n + p^3) \sim \mathcal{O}(p^2 n)$. □

Computing OLS estimate using matrix inversion is possible, however in most cases not used in practice because of the low numerical stability computing matrix inversion.

2.1.2 Computation using the Cholesky decomposition

Definition 14. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be an symmetric positive definite matrix. Then it is possible to find lower triangular matrix L so that

$$\boldsymbol{A} = \boldsymbol{L L^T} \tag{2.4}$$

We call this decomposition as *Cholesky factorization* (sometimes also *Cholesky decomposition*)

If we look at our problem of finding a solution to

$$\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} = \boldsymbol{X}^T \boldsymbol{y}, \tag{2.5}$$

we can easily rewrite it as

$$\boldsymbol{Z} \boldsymbol{w} = \boldsymbol{b} \tag{2.6}$$

where $\boldsymbol{Z} = \boldsymbol{X}^T \boldsymbol{X}$ is symmetric positive definite matrix and $\boldsymbol{b} = \boldsymbol{X}^T \boldsymbol{y}$. Because $\boldsymbol{Z}$ can be factorized as $\boldsymbol{L L^T}$ the solution can be found easily by substitution

$$\boldsymbol{L d} = \boldsymbol{b} \tag{2.7}$$

$$\boldsymbol{L}^T \boldsymbol{w} = \boldsymbol{d}, \tag{2.8}$$

where $\boldsymbol{d}$ is solved by forward substitution and $\boldsymbol{w}$ is then solved by backward substitution which ten represents $\hat{\boldsymbol{w}}^{(OLS,n)}$ estimate.

So the algorithm for solving OLS using Cholesky factorization goes as follows:

1. Compute $\boldsymbol{X}^T \boldsymbol{X}$ and $\boldsymbol{X}^T \boldsymbol{y}$.

2. Compute Cholesky factorization $\boldsymbol{Z} = \boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X}$.
3. Solve lower triangular system $\boldsymbol{L}\boldsymbol{d} = \boldsymbol{b}$ where $\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}$ for $\boldsymbol{d}$ using forward substitution.
4. Solve upper triangular system $\boldsymbol{L}^T\boldsymbol{w} = \boldsymbol{d}$ for $\boldsymbol{w}$.

Observation 15. The time complexity of solving OLS using Cholesky factorization is $\mathcal{O}(p^2 n)$

Proof. First step requires two matrix multiplication $\boldsymbol{X}^T\boldsymbol{X}$ which is $\mathcal{O}(np^2)$ and $\boldsymbol{X}^T\boldsymbol{y}$ which is $\mathcal{O}(np)$. Second step represents computing Cholesky factorization. This can be done in $\mathcal{O}(\frac{1}{3}p^3)$ [14] In the third and fourth step, we are solving triangular systems both of them require $\mathcal{O}(\frac{1}{2}p^2)$, so together $\mathcal{O}(p^2)$.

Putting all steps together we get

$$\mathcal{O}(np^2 + np + \frac{1}{3}p^3 + p^2) \tag{2.9}$$

and because we assume $n \geq p$ and most usually $n \gg p$ then multiplication $\boldsymbol{X}^T\boldsymbol{X}$ asymptotically dominates over the rest so we get $\mathcal{O}(p^2 n)$. □

Computation OLS estimate using the Cholesky factorization is more numerically stable than using matrix inversion and time complexity is asymptotically similar. This approach is however not usually used in practice as well.

2.1.3 Computation using the QR decomposition

Let us now look on a similar method of computing OLS estimate and which does not require multiplying $\boldsymbol{X}^T\boldsymbol{X}$.

Definition 16. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be any square matrix. Then it is possible to find matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ so that

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}, \tag{2.10}$$

where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is orthogonal matrix and $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ is upper triangular matrix. This decomposition is known as *QR decomposition*.

If matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ thus is not square, then decomposition can be found as

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} = \boldsymbol{Q}\begin{bmatrix}\boldsymbol{R_1}\\ \boldsymbol{0}\end{bmatrix} = \begin{bmatrix}\boldsymbol{Q_1} & \boldsymbol{Q_2}\end{bmatrix}\begin{bmatrix}\boldsymbol{R_1}\\ \boldsymbol{0}\end{bmatrix} = \boldsymbol{Q_1}\boldsymbol{R_1} \tag{2.11}$$

Where $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ is orthogonal matrix, $\boldsymbol{R_1} \in \mathbb{R}^{n\times n}$ is upper triangular matrix, $\boldsymbol{0} \in \mathbb{R}^{(m-n)\times n}$ is zero matrix, $\boldsymbol{Q_1} \in \mathbb{R}^{n\times n}$ is matrix with orthogonal columns and $\boldsymbol{Q_2} \in \mathbb{R}^{(m-n)\times n}$ is also matrix with orthogonal columns.

That means that $\boldsymbol{X}$ can be factorized as

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}. \tag{2.12}$$

Then

$$\boldsymbol{X}^T\boldsymbol{X} = (\boldsymbol{Q}\boldsymbol{R})^T\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R} \tag{2.13}$$

and because $\boldsymbol{Q}$ is orthogonal, then $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$ so that

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{R}^T\boldsymbol{R}. \tag{2.14}$$

Because $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is not square matrix then as we shown $\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R_1} \\ \boldsymbol{0} \end{bmatrix}$ so that

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{R_1}^T\boldsymbol{R_1}. \tag{2.15}$$

Where $\boldsymbol{R_1^T}$ is lower triangular matrix.

Solution to

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y} \tag{2.16}$$

is then given by

$$\boldsymbol{R_1}^T\boldsymbol{R_1}\boldsymbol{w} = \boldsymbol{R_1}^T\boldsymbol{Q_1}^T\boldsymbol{y} \tag{2.17}$$

and because we assume $\boldsymbol{X}$ have full column rank, thus is invertible, we can simplify it to

$$\boldsymbol{R_1}\boldsymbol{w} = \boldsymbol{b_1} \tag{2.18}$$

where $\boldsymbol{b_1} = \boldsymbol{Q_1}^T\boldsymbol{y}$. Because $\boldsymbol{R_1}$ is upper triangular matrix, solution of $\boldsymbol{w}$ is trivial using backward substitution. $\boldsymbol{w}$ then represents our OLS estimate $\hat{\boldsymbol{w}}^{(OLS,n)}$.

Note 17. We can see that Cholesky factorization $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{L}\boldsymbol{L}^T$ is closely connected to the QR decomposition $\boldsymbol{X} = \boldsymbol{Q_1}\boldsymbol{R_1}$ so that

$$\boldsymbol{R_1^T} = \boldsymbol{L} \tag{2.19}$$

So the algorithm for solving OLS using QR decomposition can go as follows:

1. Calculate QR decomposition $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}^T = \boldsymbol{Q_1}\boldsymbol{R_1}^T$.
2. Calculate $\boldsymbol{b_1} = \boldsymbol{Q_1}^T\boldsymbol{y}$.
3. Solve upper triangular system $\boldsymbol{R_1}\boldsymbol{w} = \boldsymbol{b_1}$ for $\boldsymbol{w}$.

QR factorization can be calculated in multiple ways. The most basic method is applying the Gram-Schmidt process to columns of matrix $\boldsymbol{X}$. This approach is not numerically stable so in practice is not used as much as two other methods. Those are *Householder transformations* and *Givens rotations*.

The time complexity of both algorithms is similar. QR decomposition of matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is

$$\mathcal{O}(2p^2n - \frac{2}{3}p^3) \sim O(p^2n). \tag{2.20}$$

On the other hand, using Givens rotation for QR decomposition of the same matrix is find citation

$$\mathcal{O}(3np^2 - p^3) \sim\sim O(p^2n). \tag{2.21}$$

so although Householder transformations are about 50% faster, both have asymptotically same time complexity. Moreover Givens rotations are more numerically stable. We speak about Givens rotations in detail in section 2.1.3 where we also make proof of the (2.21). Moreover, Givens rotations and are suitable for sparse matrices. This property of Givens rotations we use in section 2.4.4.

For now, let us look at time complexity of solving OLS using Householder transformations.

Observation 18. The time complexity of solving OLS estimate using QR decomposition by Householder transformations is $\mathcal{O}(2p^2n - \frac{2}{3}p^3) \sim \mathcal{O}(p^2n)$

Proof. In the first step, we calculate QR decomposition. Householder transformations can be done in $\mathcal{O}(2np^2 - \frac{2}{3}p^3)$. Second step consist of matrix multiplication $\boldsymbol{Q_1}^T\boldsymbol{y}$ which is $\mathcal{O}(np)$. In the last step we are solving upper triangular systems which is $\mathcal{O}(\frac{1}{2}p^2)$. Putting all steps together we get

$$\mathcal{O}(2np^2 - \frac{2}{3}p^3 + np + \frac{1}{2}p^2) \tag{2.22}$$

We can see that p^2n asymptotically dominates (this is the same in case we use Givens rotations). □

The QR decomposition is considered as a standard way of computing OLS estimate because of its high numerical stability. Let us mention that a little slower, but a more stable method of computing OLS estimate exists, and that is the singular value decomposition (SVD). On the other hand, the QR decomposition is sufficiently stable for most cases so describing SVD decomposition is out of the scope of this work.

Givens Rotation

In this section, we describe Givens rotations in detail. Moreover, we also show how to update QR decomposition in terms of including and excluding row from factorized matrix $\boldsymbol{X}$ which we use in the algorithm described in section 2.4.4.

We compute the QR factorization of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ which has full column rank so that we apply orthogonal transformation by a matrix $\boldsymbol{Q}^T$ as [15]

$$\boldsymbol{Q}^T \boldsymbol{A} = \boldsymbol{R} \tag{2.23}$$

where $\boldsymbol{Q}$ is product of orthogonal matrices. One such example of orthogonal matrix can be:

$$\boldsymbol{Q} = \begin{bmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{bmatrix} \tag{2.24}$$

This matrix is indeed orthogonal since

$$\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I} \tag{2.25}$$

Moreover, if we multiply this orthogonal matrix with some column vector $\boldsymbol{x} \in \mathbb{R}^{2\times 1}$ thus $\boldsymbol{Q}\boldsymbol{x}$; as a result, we get a vector of the same length but which is rotated clockwise by φ radians. When we say that vector has the same length we mean that L^2 norm of such vector stays the same. This is an important property of all orthogonal matrices (operations that preserve L^2 norm are known as *unitary transformations*). Let us verify this claim. Let us have an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ and any column vector $\boldsymbol{x} \in \mathbb{R}^{m\times 1}$ then

$$\|\boldsymbol{Q}\boldsymbol{x}\|^2 = (\boldsymbol{Q}\boldsymbol{x})^T(\boldsymbol{Q}\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{x} = \|\boldsymbol{x}\|^2 . \tag{2.26}$$

So as we said, the idea is to create a series of orthogonal matrices which gradually rotates column vectors of $\boldsymbol{A}$, so we obtain zeros under diagonal. One such method - *Givens rotation* uses *Givens matricies* that zero one element under diagonal at a time.

The example of (2.24) is not random. Let us look at this orthogonal matrix one more time and let us multiply this matrix with some vector.

$$\begin{bmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} r \\ z \end{bmatrix} \tag{2.27}$$

To introduce this zeroing effect, we need to rotate this vector so that it is parallel to $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$. So we only need to find $\cos(\varphi)a + \sin(\varphi) = r$. As we know rotation with $\boldsymbol{Q}$ preserve L^2 norm. That means if we want to $z = 0$ then r must be equal to L^2 norm of the vector $\begin{bmatrix} a \\ b \end{bmatrix}$; thus $r = \sqrt{a^2 + b^2}$. Then the solution is trivial. We do not even need to calculate φ because if we put

$$\cos(\varphi) = \frac{a}{\sqrt{a^2 + b^2}} \tag{2.28}$$

and

$$\sin(\varphi) = \frac{b}{\sqrt{a^2 + b^2}} \tag{2.29}$$

then

$$r = \frac{a^2}{\sqrt{a^2 + b^2}} + \frac{b^2}{\sqrt{a^2 + b^2}} = \sqrt{a^2 + b^2} \tag{2.30}$$

$$z = \frac{-ab}{\sqrt{a^2 + b^2}} + \frac{ab}{\sqrt{a^2 + b^2}} = 0. \tag{2.31}$$

Practically used algorithm of computing $\cos\varphi$ and $\sin\varphi$ is slightly different because we want to prevent overflow. This algorithm is described by the following pseudocode:

Algorithm 1: Rotate

Input: a, b
Output: cos, sin
$sin \leftarrow \emptyset$;
$cos \leftarrow \emptyset$;
if $b == 0$ **then**
  $sin \leftarrow 0$;
  $cos \leftarrow 1$;
else if $abs(b) \geq abs(a)$ **then**
  $cotg \leftarrow \frac{a}{b}$;
  $sin \leftarrow \frac{1}{\sqrt{1+(cotg)^2}}$;
  $cos \leftarrow sin\ cotg$;
else
  $tan \leftarrow \frac{b}{b}$;
  $cos \leftarrow \frac{1}{\sqrt{1+(tan)^2}}$;
  $sin \leftarrow cos\ tan$;
end
return $cos\ ,\ sin$;

We can scale this same idea to higher dimensions. Let us denote matrix $\boldsymbol{Q}(i,j) \in \mathbb{R}^{m \times m}$ defined as

$$\boldsymbol{Q}(i,j) = \begin{array}{c} \\ \\ i \\ \\ j \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & i & & j & & \end{array} \\ \begin{bmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & c & \dots & s & \dots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \dots & -s & \dots & c & \dots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{bmatrix} \end{array}$$

where $c = \cos(\varphi)$ and $s = \sin(\varphi)$ for some φ. That means this matrix is orthogonal. Now if we have some column vector $\boldsymbol{x} \in \mathbb{R}^{m\times 1}$ and we calculate c and s by means of 2.28 and 2.29 for $a := x_i$ and $b := x_j$. Finally if we multiply $\boldsymbol{Q}(i,j)\boldsymbol{x} = \boldsymbol{p}$ we can see that $\boldsymbol{p}$ is same as vector $\boldsymbol{x}$ except p_i and p_j so that:

$$\boldsymbol{Q}(i,j)\boldsymbol{x} = \boldsymbol{Q}(i,j) \begin{matrix} \\ \\ i \\ \\ j \\ \\ \end{matrix} \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_j \\ \vdots \\ x_m \end{bmatrix} = \boldsymbol{p} = \begin{matrix} \\ \\ i \\ \\ j \\ \\ \end{matrix} \begin{bmatrix} x_1 \\ \vdots \\ r \\ \vdots \\ 0 \\ \vdots \\ x_m \end{bmatrix}$$

where $r = \sqrt{x_i^2 + x_j^2}$. If we have matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ instead of only one column $\boldsymbol{x}$ it works the same way. So if we have such matrix we need to create matrix $\boldsymbol{Q}(i,j)$ for each a_{ij} under the diagonal in order to create upper triangular matrix. Usually we are zeroing columns so that we start with a_{12} then $a_{13} \dots a_{1_n}$. Then we start with second column a_{23} and so on. That means we need to create in total $e = \frac{p^2-p}{2} + np - p^2$ matrices $\boldsymbol{Q}(i,j)$. We can denote this sequence of matrices as $\boldsymbol{Q_1}, \boldsymbol{Q_2}, \dots \boldsymbol{Q_e}$. The QR decomposition then looks like

$$\boldsymbol{Q_e} \dots \boldsymbol{Q_2}\boldsymbol{Q_1}\boldsymbol{A} = \boldsymbol{R} \tag{2.32}$$

where $\boldsymbol{Q_e} \dots \boldsymbol{Q_2}\boldsymbol{Q_1}$ is actually $\boldsymbol{Q}^T$. So $\boldsymbol{Q}$ is obtained by

$$\boldsymbol{Q} = \boldsymbol{Q_1^T}\boldsymbol{Q_2^T} \dots \boldsymbol{Q_e^T}. \tag{2.33}$$

We can see that for each row of the matrix, we need one additional matrix $\boldsymbol{Q_i}$, and with such matrix, we are multiplying only two rows. Moreover, the rows are getting shorter after each finished column. So the total number of operations can be expressed as

$$\sum_{i=1}^{n}\sum_{j=i+1}^{p} 6(n-i+1) \approx 6np^2 - 3np^2 - 3p^3 + 2p^3 = 3np^2 - p^3. \tag{2.34}$$

2.2 FAST-LTS

In this section, we introduce the FAST-LTS algorithm[16]. It is, as well as other algorithms we introduce, iterative algorithm. We discuss all main components of the algorithm starting with its core idea called a concentration step which authors call a *C-step*.

2.2.1 C-step

We show that from an existing LTS estimate $\hat{\boldsymbol{w}}$ we can construct a new LTS estimate $\hat{\boldsymbol{w}}_{\boldsymbol{new}}$ so that value the objective function at $\hat{\boldsymbol{w}}_{\boldsymbol{new}}$ is less or equal to the value at $\hat{\boldsymbol{w}}$. Based on this property, the algorithm creates a sequence of LTS estimates which leading to better results.

Theorem 19. Consider $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$. Let us also have $\boldsymbol{w_0} \in \mathbb{R}^p$ and $\boldsymbol{m_0} \in Q^{(n,h)}$. Let us put $L_0 = \sum_{i=1}^{n} m_i^{(0)} r_i^2(\boldsymbol{w_0})$ where $r_i^2(\boldsymbol{w_0}) = y_i - (w_1^0 x_1^i + w_2^0 x_2^i + \ldots + w_p^0 x_p^i)$. Let us take $\hat{n} = \{1, 2, \ldots, n\}$ and mark $\pi : \hat{n} \to \hat{n}$ permutation of $\hat{n}$ such that $|r_0(\pi(1))| \leq |r_0(\pi(2))| \leq \ldots \leq |r_0(\pi(n))|$ and mark $\boldsymbol{m_1} \in Q^{(n,h)}$ such that $m_i^1 = 1$ for $i \in \{\pi(1), \pi(2), ...\pi(h)\}$ and $m_i^1 = 0$ otherwise. This means that $\boldsymbol{m_1}$ corresponds to h subset with smallest absolute residuals $r_i^2(\boldsymbol{w_0})$.

Finally, take $\mathrm{OF}^{(OLS, \boldsymbol{M_1 X}, \boldsymbol{M_1 y})}(\boldsymbol{w})$ least squares fit on m_1 subset of observations and its corresponding $L_1 = \sum_{i=1}^{n} m_i^{(1)} r_i^2(\boldsymbol{w_1})$. Then

$$L_1 \leq L_0 \tag{2.35}$$

Proof. Because we take h observations with smallest absolute residuals $r_i^2(\boldsymbol{w_0})$, then for sure $\sum_{i=1}^{n} m_i^{(1)} r_i^2(\boldsymbol{w_0}) \leq \sum_{i=1}^{n} m_i^{(0)} r_i^2(\boldsymbol{w_0}) = L_0$. When we take into account that the OLS estimate minimizes the objective function of $\boldsymbol{m_1}$ subset of observations, then for sure $L_1 = \sum_{i=1}^{n} m_i^{(1)} r_i^2(\boldsymbol{w_1}) \leq \sum_{i=1}^{n} m_i^{(1)} r_i^2(\boldsymbol{w_0})$.

Together we get $L_1 = \sum_{i=1}^{n} m_i^{(1)} r_i^2(\boldsymbol{w_1}) \leq \sum_{i=1}^{n} m_i^{(1)} r_i^2(\boldsymbol{w_0}) \leq \sum_{i=1}^{n} m_i^{(0)} r_i^2(\boldsymbol{w_0}) = L_0$ □

Corollary 20. Based on the previous theorem, using some $\hat{\boldsymbol{w}}_{\boldsymbol{old}}^{\boldsymbol{OLS(h,n)}}$ of $\boldsymbol{m_{old}}$ subset of observations we can construct $\boldsymbol{m_{new}}$ subset with corresponding $\hat{\boldsymbol{w}}_{\boldsymbol{new}}^{\boldsymbol{OLS(H_{new})}}$ such that $L_{new} \leq L_{old}$. Applying the theorem repetitively leads to the iterative sequence of $L_1 \leq L_2 \leq \ldots$. One step, called *C-step* of this process is described by the following pseudocode. Note that for C-step we need only $\hat{\boldsymbol{w}}$ without the need of passing $\boldsymbol{m}$.

Algorithm 2: C-step

Input: dataset consiting of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$, $\hat{\boldsymbol{w}}_{\boldsymbol{old}} \in \mathbb{R}^{p \times 1}$
Output: $\hat{\boldsymbol{w}}_{\boldsymbol{new}}$, m_{new}

```
R ← ∅;
for i ← 1 to n do
    R ← R ∪ {|y_i − ŵ_old x_i^T|};
end
m_new ← select set of h smallest absolute residuals from R;
ŵ_new ← OLS fit on m_new subset; return ŵ_new, m_new;
```

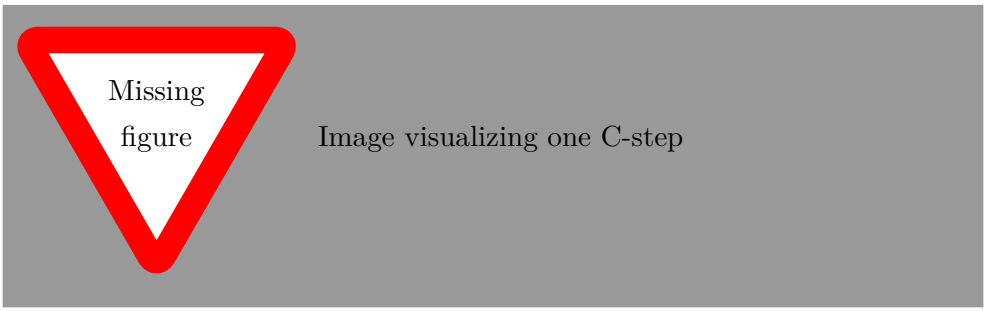


Figure 2.1: todo caption

One step of algorithm C-step is visualized on figure 2.1.

Observation 21. The time complexity of algorithm C-step 2 is asymptotically similar to the time complexity as OLS fit. Thus $O(p^2n)$.

Proof. In C-step we must compute n absolute residuals. Computation of one absolute residual consists of matrix multiplication of shapes $1 \times p$ and $p \times 1$ that gives us $\mathcal{O}(p)$. So time of computation n residuals is $\mathcal{O}(np)$. Next, we must select a set of h smallest residuals which can be done in $\mathcal{O}(n)$ using a modification of algorithm QuickSelect [17]. Finally, we must compute $\hat{w}$ OLS estimate on a h subset of data. Because h is linearly dependent on n, we can say that it is $\mathcal{O}(p^2n + p^3)$ which is asymptotically dominant against previous steps which are $\mathcal{O}(np + n)$. Because we assume$n \geq p$ then $\mathcal{O}(p^2n + p^3) \sim \mathcal{O}(p^2n)$ □

As we stated above, repeating C-step leads to sequence of $\hat{\boldsymbol{w}}_\mathbf{1}, \hat{\boldsymbol{w}}_\mathbf{2} \ldots$ on subsets $\boldsymbol{m}_\mathbf{1}, \boldsymbol{m}_\mathbf{2} \ldots$ with corresponding $L_1 \geq L_2 \geq \ldots$. One could ask if this sequence converges, so that $L_i == L_{i+1}$. Answer to this question is presented by the following theorem.

Theorem 22. Sequence of C-step converges to $\hat{\boldsymbol{w}}_\boldsymbol{k}$ after maximum of $k = \binom{n}{h}$ so that

$$L_k = L_i \,, \forall i \geq k. \tag{2.36}$$

Proof. Since $\mathrm{OF}_\mathrm{D}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_\boldsymbol{i})$ is non-negative and $\mathrm{OF}_\mathrm{D}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_\boldsymbol{i}) \leq \mathrm{OF}_\mathrm{D}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_{\boldsymbol{i+i}})$ the sequence converges. $\hat{\boldsymbol{w}}_\boldsymbol{i}$ is computed out of subset $\boldsymbol{m}_\boldsymbol{i} \in Q^{(n,h)}$. Since $Q^{(n,h)}$ is finite, namely its size is $\binom{n}{h}$, the sequence converges at the latest after this number of steps. □

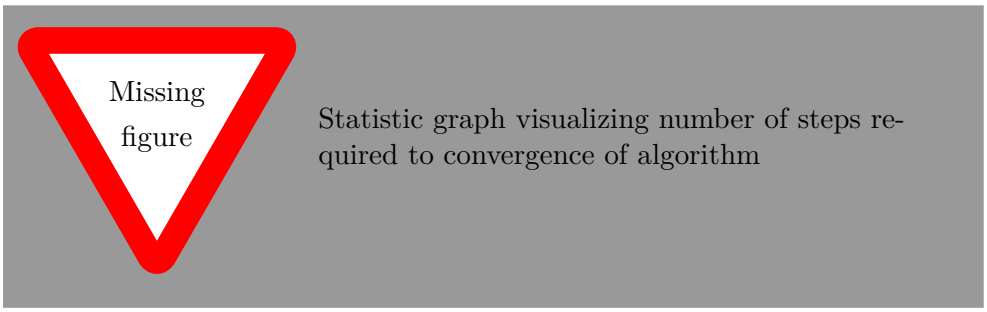
The theorem gives us a clue to create algorithm described by the following pseudocode.

Algorithm 3: Repeat-C-step

Input: dataset consiting of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$, $\hat{\boldsymbol{w}}_{\boldsymbol{old}} \in \mathbb{R}^{p \times 1}$, $\boldsymbol{m_0}$
Output: $\hat{\boldsymbol{w}}_{\boldsymbol{final}}$, $\boldsymbol{m}_{\boldsymbol{final}}$

$\hat{\boldsymbol{w}}_{\boldsymbol{new}} \leftarrow \emptyset$;
$\boldsymbol{m}_{\boldsymbol{new}} \leftarrow \emptyset$;
$L_{new} \leftarrow \infty$;
while $True$ **do**
  $L_{old} \leftarrow \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_{\boldsymbol{old}})$;
  $\hat{\boldsymbol{w}}_{\boldsymbol{new}}$, $H_{new} \leftarrow$ C-step($\boldsymbol{X}$, $\boldsymbol{y}$, $\hat{\boldsymbol{w}}_{\boldsymbol{old}}$);
  $L_{new} \leftarrow \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_{\boldsymbol{new}})$;
  if $Lold == Lnew$ **then**
    break
  end
  $\hat{\boldsymbol{w}}_{\boldsymbol{old}} \leftarrow \hat{\boldsymbol{w}}_{\boldsymbol{new}}$
end
return $\hat{\boldsymbol{w}}_{\boldsymbol{new}}$, $\boldsymbol{m}_{\boldsymbol{new}}$;

It is important to note, that although the maximum number of steps of this algorithm is $\binom{n}{h}$ in practice, it is shallow, most often under 20 steps as can be seen on figure 2.2.



Statistic graph visualizing number of steps required to convergence of algorithm

Figure 2.2: todo caption

That is not enough for the algorithm Repeat-C-step to converge to the global minimum. That gives us an idea of how to create the final algorithm. [16]

Choose a lot of initial subsets $\boldsymbol{m_1}$ and on each of them apply algorithm Repeat-C-step. From all converged subsets with corresponding $\hat{\boldsymbol{w}}$ estimates choose the one with the lowest value of $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\hat{w})$.

Before we can construct final the algorithm, we must decide how to choose initial subset $\boldsymbol{m_1}$ and how many of them mean "*a lot*". First, let us focus on how to choose an initial subset $\boldsymbol{m_1}$.

2.2.2 Choosing initial $m_1$subset

It is important to note, that when we choose $\boldsymbol{m_1}$ subset such that it contains outliers, then iteration of C-steps usually does not converge to good results, so we should focus on methods with non zero probability of selecting $\boldsymbol{m_1}$ such that it does not contain outliers. There are many possibilities of how to create an initial $\boldsymbol{m_1}$ subset. Let us start with the most trivial one.

Random selection

Most basic way of creating $\boldsymbol{m_1}$ subset is simply to choose random $\boldsymbol{m_1} \in Q^{(n,h)}$. The following observation shows that it is not the best way.

Observation 23. With an increasing number of data samples, thus with increasing n, the probability of choosing among k random selections of $\boldsymbol{m_{1_1}}, \ldots, \boldsymbol{m_{1_k}}$ the probability of selecting at least one $\boldsymbol{m_{1_i}}$ such that its corresponding data samples does not contains outliers, goes to 0.

Proof. Consider dataset of n containing $\epsilon > 0$ relative amount of outliers. Let h be chosen conservatively so that $h = [(n/2] + [(p+1)/2]$ and k is the number of selections random h subsets. Then

$$\begin{aligned} P(\textit{one random data sample not outliers}) &= (1-\epsilon) \\ P(\textit{one subset without outliers}) &= (1-\epsilon)^h \\ P(\textit{one subset with at least one outlier}) &= 1-(1-\epsilon)^h \\ P(\textit{m subsets with at least one outlier in each}) &= (1-(1-\epsilon)^h)^k \\ P(\textit{m subsets with at least one subset without outliers}) &= 1-(1-(1-\epsilon)^h)^k \end{aligned}$$

Because $n \to \infty$, then $(1-\epsilon)^h \to 0$, then $1-(1-\epsilon)^h \to 1$, then $(1-(1-\epsilon)^h)^m \to 1$, then $1-(1-(1-\epsilon)^h)^m \to 0$ □

That means that we should consider other options for selecting $\boldsymbol{m_1}$ subset. If we would like to continue with selecting some random subsets, previous observation gives us a clue, that we should choose it independent of n. Authors of the algorithm came with such a solution, and it goes as follows.

P-subset selection

Let us choose vector $\boldsymbol{c} \in Q^{(n,p)}$. Next we compute rank of matrix $\boldsymbol{X}_C = \boldsymbol{C}\boldsymbol{X}$, where $C = diag(c)$. If $rank(\boldsymbol{X}_C) < p$ add randomly selected rows to $\boldsymbol{X}_C$ without repetition until $rank(\boldsymbol{X}_C) = p$. Let us from now on suppose that $rank(\boldsymbol{X}_C) = p$. Next let us mark $\hat{\boldsymbol{w}}_{\boldsymbol{0}} = \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(c)$ and corresponding $r_1(\hat{\boldsymbol{w}}_{\boldsymbol{0}}), r_2(\hat{\boldsymbol{w}}_{\boldsymbol{0}}), \ldots, r_n(\hat{\boldsymbol{w}}_{\boldsymbol{0}})$ residuals. Now mark $\hat{n} = \{1, 2, \ldots, n\}$ and let $\pi : \hat{n} \to \hat{n}$ be permutation of $\hat{n}$ such that $|r(\pi(1))| \leq |r(\pi(2))| \leq \ldots \leq |r(\pi(n))|$.

Finally, mark $\boldsymbol{m_1} \in Q^{(n,h)}$ such that $m_i^1 = 1$ for $i \in \{\pi(1), \pi(2), ...\pi(h)\}$ and $m_i^1 = 0$ otherwise.

Observation 24. With the increasing number of data samples, thus with increasing n, the probability of choosing among k random selections of $\boldsymbol{c_{1_1}}, \ldots, \boldsymbol{c_{1_k}}$ the probability of selecting at least one $\boldsymbol{c_{1_i}}$ such that its corresponding data samples do not contains outliers, goes to

$$1 - (1 - (1 - \epsilon)^h)^k > 0. \tag{2.37}$$

Proof. Similarly as in previous observation. □

Last missing piece of the algorithm is determining the number of k initial $\boldsymbol{m_1}$ subsets, which maximize the probability to at least one of them converges to correct solution. Simply put, the more, the better. So before we answer this question accurately, let us discuss some key observations about the algorithm.

2.2.3 Speed-up of the algorithm

In this section, we describe essential observations which help us to formulate the final algorithm. In two subsections we describe how to optimize the current algorithm.

Selective iteration

The most computationally demanding part of one C-step is computation of OLS fit on $\boldsymbol{m_i}$ subset and then the calculation of n absolute residuals. How we stated above, convergence is usually achieved under 20 steps. So for fast algorithm run, we would like to repeat C-step as little as possible and at the same time do not lose the performance of the algorithm.

Due to that convergence of repeating C-step is very fast, it turns out, that we can distinguish between starts that leads to good solutions and those which does not, even after a small amount of the C-steps iterations. Based on empiric observation, we can distinguish good or bad solution already after two or three iterations of C-steps based on $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_{\boldsymbol{3}})$ or $ofl\hat{t}s(\hat{\boldsymbol{w}}_{\boldsymbol{4}})$ respectively.

So even though authors do not specify the size of k explicitly, they propose that after a few C-steps we can choose (say 10) best solutions among all $\boldsymbol{m_1}$ starting subsets and continue iteration of the C-steps till convergence only on those best solutions. Authors refer to this process as to *selective iteration*.

Nested extension

C-step computation is usually very fast for small n. Problem starts with very high n say $n > 10^3$ because we need to compute OLS on $\boldsymbol{m_i}$ subset of size h which is dependent on n. And then calculate n absolute residuals.

Authors came up with a solution they call *nested extension*. We describe it briefly now.

- If n is greater than limit l, we create subset of data samples $L\,, |L| = l$ and divide this subset into s disjunctive sets $P_1, P_2, \ldots, P_s\,, |P_i| = \frac{l}{s}\,, P_i \cap P_j = \emptyset\,, \bigcup_{i=1}^{s} P_i = L$.

- For every P_i we set number of starts $m_{P_i} = \frac{m}{l}$.

- Next in every P_i we create m_{P_i} number of initial $H_{P_{i_1}}$ subsets and iterate C-steps for two iterations.

- Then we choose 10 best results from each subsets and merge them together. We get family of sets F_{merged} containing 10 best $H_{P_{i_3}}$ subsets from each P_i.

- On each subset from F_{merged} family of subsets we again iterate 2 C-steps and then choose 10 best results.

- Finally we use these best 10 subsets and use them to iterate C-steps till convergence.

- As a result we choose best of those 10 converged results.

2.2.4 Putting all together

We described all major parts of the algorithm FAST-LTS. One last thing we need to mention is that even though C-steps iteration usually converges under 20 steps, it is appropriate to introduce two parameters i_{max} and t which limits the number of C-steps iterations in some rare cases when convergence is too slow. Parameter i_{max} denotes the maximum number of iterations in final C-step iteration till convergence. Parameter t denotes threshold stopping criterion such that $|\operatorname{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_{\boldsymbol{i}}) - \operatorname{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_{\boldsymbol{i+1}})| \leq t$ instead of $\operatorname{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_{\boldsymbol{i}}) = \operatorname{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_{\boldsymbol{i+1}})$. When we put all together, we get *FAST-LTS* algorithm which

is described by the following pseudocode.

Algorithm 4: FAST-LTS

Input: $\boldsymbol{X} \in \mathbb{R}^{n \times p}, \boldsymbol{y} \in \mathbb{R}^{n \times 1}, m, l, s, i_{max}, t$
Output: $\hat{\boldsymbol{w}}_{\boldsymbol{final}}, \boldsymbol{m}_{\boldsymbol{final}}$

$\hat{\boldsymbol{w}}_{\boldsymbol{final}} \leftarrow \emptyset$;
$\boldsymbol{m}_{\boldsymbol{final}} \leftarrow \emptyset$;
$F_{best} \leftarrow \emptyset$;
if $n \geq l$ **then**
  $F_{merged} \leftarrow \emptyset$;
  for $i \leftarrow 0$ ***to*** s **do**
    $F_{selected} \leftarrow \emptyset$;
    for $j \leftarrow 0$ ***to*** $\frac{l}{s}$ **do**
      $F_{initial} \leftarrow Selective\ iteration(\frac{m}{l})$;
      for $\boldsymbol{m_i}$ ***in*** $F_{initial}$ **do**
        $\boldsymbol{m_i} \leftarrow Iterate\ C\ step\ few\ times(\boldsymbol{m_i})$;
        $F_{selected} \leftarrow F_{selected} \cup \{\boldsymbol{m_i}\}$;
      end
    end
    $F_{merged} \leftarrow F_{merged} \cup Select\ 10\ best\ subsets\ from\ F_{selected}$;
  end
  for $\boldsymbol{m_i}$ ***in*** F_{merged} **do**
    $\boldsymbol{m_i} \leftarrow Iterate\ C\ step\ few\ times(\boldsymbol{m_i})$;
    $F_{best} \leftarrow F_{best} \cup \{\boldsymbol{m_i}\}$;
  end
  $F_{best} \leftarrow Select\ 10\ best\ subsets\ from\ F_{best}$;
else
  $F_{initial} \leftarrow Selective\ iteration(m)$;
  $F_{best} \leftarrow Select\ 10\ best\ subsets\ from\ F_{initial}$;
end
$F_{final} \leftarrow \emptyset$;
$W_{final} \leftarrow \emptyset$;
for $\boldsymbol{m_i}$ ***in*** F_{best} **do**
  $\boldsymbol{m_i}, \hat{\boldsymbol{w}}_{\boldsymbol{i}} \leftarrow Iterate\ C\ step\ till\ convergence(\boldsymbol{m_i}, i_{max}, t)$;
  $F_{final} \leftarrow F_{final} \cup \{\boldsymbol{m_i}\}$;
  $W_{final} \leftarrow W_{final} \cup \{\hat{\boldsymbol{w}}_{\boldsymbol{i}}\}$;
end
$\hat{\boldsymbol{w}}_{\boldsymbol{final}}, \boldsymbol{m}_{\boldsymbol{final}} \leftarrow$
  $select\ \hat{\boldsymbol{w}}_{\boldsymbol{i}}\ and\ \boldsymbol{m_i}\ with\ smallest\ \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\hat{\boldsymbol{w}}_{\boldsymbol{i}})\ from\ F_{final}$;
return $\hat{\boldsymbol{w}}_{\boldsymbol{final}}, \boldsymbol{m}_{\boldsymbol{final}}$;

2.3 Feasible solution

In this section we introduce feasible solution algorithm from [6]. It is based on the strong necessary condition described at definition 10. The basic idea can be described as follows.

Let us consider that we have some $\boldsymbol{m} \in Q^{(n,h)}$ and denote $O_m = \{i \in \{1, 2, \ldots, n\}; w_i = 1\}$ and $Z_m = \{j \in \{1, 2, \ldots, n\}; w_j = 0\}$ thus sets of indexes of positions where is 0 respectively 1 in vector $\boldsymbol{m}$. We can think about it as indexes of observations in h set and g set respectively. Then we can mark $\boldsymbol{m}^{(i,j)}$ as a vector which is constructed by swap of its ith and jth element where $i \in O$ and $j \in Z$. Such vector corresponds to vector $\boldsymbol{m_{swap}}$ which we marked at definition 10.

With this in mind, let us mark

$$\Delta S_{i,j}^{(\boldsymbol{m})} = \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}^{(i,j)}) - \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}) \tag{2.38}$$

thus a change of the LTS objective function by swapping one observation from not trimmed subset with another from trimmed subset. To calculate this we can obviously first calculate the $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m})$ and the $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}^{(i,j)})$ and finally subtract both results. Although it is a option, it it computationally exhaustive. So the question is if there is easier way of calculating $\Delta S_{i,j}^{(\boldsymbol{m})}$. The answer is positive and we describe it now.

Let us mark $M = diag(\boldsymbol{m})$ and $M^{(i,j)} = diag(\boldsymbol{m}^{(i,j)})$ and also $\boldsymbol{Z_M} = (\boldsymbol{X}^T \boldsymbol{M}^T \boldsymbol{X})$ For now let us assume that we also have $\boldsymbol{Z_M}^{-1}$ and $\hat{\boldsymbol{w}} = \boldsymbol{Z_M}^{-1} \boldsymbol{X}^T \boldsymbol{M} \boldsymbol{y}$. We now want to calculate $\Delta S_{i,j}^{(\boldsymbol{m})}$. Let us mark vector of residuals $\boldsymbol{r}^{(\boldsymbol{m})} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{w}}$ and also $d_{r,s} = \boldsymbol{x_r} \boldsymbol{Z_M}^{-1} \boldsymbol{x_s}$ then by equation introduced in [18] we get

$$\Delta S_{i,j}^{(\boldsymbol{m})} = \frac{(r_j^{(\boldsymbol{m})})^2(1 - d_{i,i}) - (r_i^{(\boldsymbol{m})})^2(1 + d_{j,j}) + 2r_i^{(\boldsymbol{m})} r_j^{(\boldsymbol{m})} d_{j,j}}{(1 - d_{i,i})(1 + d_{j,j}) + d_{i,j}^2}. \tag{2.39}$$

Let us now describe the core of the algorithm. It is similar to the FAST-LTS algorithm in the sense of iterative refinement of a h subset. So let us assume that we have some vector $\boldsymbol{m} \in Q^{(n,h)}$. No we compute $\Delta S_{i,j}^{(\boldsymbol{m})}$ for all $i \in O$ and $j \in Z$. This may lead to several different outcomes:

1. all $S_{i,j}^{(\boldsymbol{m})}$ are non-negative
2. one $S_{i,j}^{(\boldsymbol{m})}$ is negative
3. multiple $S_{i,j}^{(\boldsymbol{m})}$ are negative

In the first case, all $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}^{(i,j)})$ are higher or the same as the $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m})$ so none swap will lead to an improvement. That also means that strong necessary condition is satisfied and the algorithm ends.

In the second and third case, the strong necessary condition is not satisfied, and we can make the swap. In the second case, it is easy which one to choose because we have only one. In the third case, we have a couple of options again:

1. use the first swap that leads to the improvement

2. from all possible swaps choose one that has highest improvement value $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}^{(i,j)})$

3. use the first swap that has improvement higher than some given threshold

In terms of complexity, all three options give the same results, because to find a feasible solution you need to evaluate all pair swaps. In practice, the third option is the winner because it leads to the least amount of iterations. On the other hand, as we said this does not improve the complexity of the algorithm, so from now on let us assume that we use the case number two. So if there negative $S_{i,j}^{(\boldsymbol{m})}$, we choose the one with the lowest value, make the swap and repeat the process.

The algorithm ends when there is no possible improvement, i.e., when all $S_{i,j}^{(\boldsymbol{m})}$ are non-negative. The number of iterations needed till algorithm stops is usually quite low, but for practical usage, it is still convenient to use some parameter i_{max} after which algorithm stops without finding h subset satisfying the strong necessary condition. One step of this algorithm we call optimal

swap additive algorithm (OSAA) is described by the following pseudocode.

Algorithm 5: OSAA

Input: $\boldsymbol{Z_M}^{-1} \in \mathbb{R}^{p\times p}$, $\boldsymbol{r^{(m)}} \in \mathbb{R}^{n\times 1}$, O_m, Z_m, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$
Output: $\hat{\boldsymbol{w}}_{\boldsymbol{new}}$, m_{new}

```
S ← 0;
i_swap ← ∅;
j_swap ← ∅;
for m_i ∈ O_m do
    for m_j ∈ Z_m do
        r_i^(m) = r_{m_i}^(m) ;
        r_j^(m) = r_{m_j}^(m) ;
        d_{i,i} = x_{m_i} Z_M^{-1} x_{m_i}^T ;
        d_{i,j} = x_{m_i} Z_M^{-1} x_{m_j}^T ;
        d_{j,j} = x_{m_j} Z_M^{-1} x_{m_j}^T ;
        S_tmp = calculate ΔS_{i,j}^(m) by (2.39) ;
        if S_tmp < S then
            S ← S_tmp;
            i_swap ← m_i;
            j_swap ← m_j;
        end
    end
end
return i_swap, j_swap, S;
```

Observation 25. The time complexity of the OSAA 5 is $\mathcal{O}(n^2p^2)$

Proof. All $d_{i,i}$ and $d_{j,j}$ can be calculated before the for loops. To calculate $d_{i,i}$ resp. $d_{j,j}$ we need co multiply vector $\in \mathbb{R}^p$ with matrix $\in \mathbb{R}^{p\times p}$ and vector $\in \mathbb{R}^p$ that is $\mathcal{O}(p^2)$. For all $d_{i,i}$ this have to be done h times and for $d_{j,j}$ $n-h$ times. So all together it is $\mathcal{O}(np^2)$.

The two loops go through all pairs; thus it is $\mathcal{O}(n^2)$. $d_{i,j}$ can be calculated in the loop, and it can be done in $\mathcal{O}(p^2)$.

If we put everything together we get $\mathcal{O}(np^2 + n^2p^2) \sim \mathcal{O}(n^2p^2)$. □

One run of this iteration process leads to some local optimum, i.e., set satisfying the strong necessary condition. In [6] they refer to to this set as to *feasible set*. This because it is not global optima the algorithm needs to be run multiple times say t times. A h subset with the smallest value of the objective function is chosen as the final solution.

Discussion about how to find the initial h subset resp. initial $\boldsymbol{m}$ was already discussed when describing FAST-LTS algorithm. More importantly, as we already suggested h subset satisfying weak necessary condition do not need to satisfy strong necessary condition so passing such a h subset as input

to this algorithm is another option and we discuss it in detail later. We now describe a feasible solution algorithm (FSA) with pseudocode, and we assume that we already have some function that generates for us h subsets, e.g., random one.

Algorithm 6: FSA

Input: $\boldsymbol{X} \in \mathbb{R}^{n \times p}, \boldsymbol{y} \in \mathbb{R}^{n \times 1}, i_{max}, t$
Output: $\hat{\boldsymbol{w}}_{\boldsymbol{final}}$, $\boldsymbol{m}_{\boldsymbol{final}}$

$\hat{\boldsymbol{w}}_{\boldsymbol{final}} \leftarrow \emptyset$;
$\boldsymbol{m}_{\boldsymbol{final}} \leftarrow \emptyset$;
$R \leftarrow \emptyset$;
for $k \leftarrow 0$ ***to*** t **do**
  $\boldsymbol{m} \leftarrow generate\_intial\_subset()$; // e.g. random $\boldsymbol{m} \in Q^{(n,h)}$
  $l \leftarrow 0$;
  while $True$ **do**
    $\boldsymbol{M} \leftarrow diag(\boldsymbol{m})$;
    $\boldsymbol{Z_M} = (\boldsymbol{X}^T \boldsymbol{M}^T \boldsymbol{X})$;
    ${\boldsymbol{Z_M}}^{-1} = calculate\ inversion\ of\ \boldsymbol{Z_M}$;
    $\hat{\boldsymbol{w}} \leftarrow \mathrm{OF}^{(OLS, \boldsymbol{MX}, \boldsymbol{My})}$;
    $\boldsymbol{r}^{(\boldsymbol{m})} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{w}}$;
    $S_{i,j}, i, j = OSAA(\ {\boldsymbol{Z_M}}^{-1}, \boldsymbol{r}^{(\boldsymbol{m})}, O_m, Z_m, \boldsymbol{X})$;
    if $S_{i,j} \geq 0$ ***or*** $l \geq i_{max}$ **then**
      $\boldsymbol{m}_{\boldsymbol{final}} \leftarrow \boldsymbol{m}$;
      $\hat{\boldsymbol{w}} \leftarrow \mathrm{OF}^{(OLS, \boldsymbol{MX}, \boldsymbol{My})}$;
      $R \leftarrow R \cup \{\boldsymbol{m}_{\boldsymbol{final}}, \hat{\boldsymbol{w}}\}$;
      break;
    end
    else
      $\boldsymbol{m} \leftarrow \boldsymbol{m}^{(i,j)}$;
    end
  end
end
$\boldsymbol{m}_{\boldsymbol{final}}, \hat{\boldsymbol{w}}_{\boldsymbol{final}} \leftarrow$ select best from R based on smallest value of $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}$;
return $\hat{\boldsymbol{w}}_{\boldsymbol{final}}$, $\boldsymbol{m}_{\boldsymbol{final}}$;

Observation 26. In the main loop beside running OSAA which time complexity is $\mathcal{O}(n^2p^2)$ we need to recalculate $\hat{\boldsymbol{w}}$ using the matrix inversion which time complexity is $\mathcal{O}(p^2n)$ (see observation 13). Main loop of the FSA runs up to l iterations for each start t. So the time complexity of whole algorithm is $\mathcal{O}(lt(n^2p^2 + p^2n))$. Because l and t are usually quite low, we can see that the FSA time complexity is dominated by the OSAA.

In this section, we've described the FSA algorithm. In the next section, we introduce a very similar algorithm, but which have higher numerical stability

and also contains various optimizations for higher performance.

2.4 OEA, MOEA, MMEA

In the FSA algorithm, we assumed that after each cycle of OSAA 5 we need to recalculate inversion $(\boldsymbol{X}^T\boldsymbol{X})$ together with $\hat{\boldsymbol{w}}$. In this section, we introduce a different approach described in [3] so that we can update it instead of recalculating. Moreover, these ideas lead to bounding condition of FSA which not only improves the speed but also gives options to construct different algorithms.

In section 2.3 we introduce additive formula 2.39 which corresponds to $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}^{(i,j)}) - \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m})$. Let us now try to obtain similar formula but with the focus on how the individual elements in our current algorithm change namely the inversion $\boldsymbol{Z}^{-1}$ and $\hat{\boldsymbol{w}}$. We also split the idea of swap the i, j into the insertion of j and remove of i. Thus we describe how individual elements are affected after adding one row and also removing one row. Moreover, during this derivation, we obtain two different approaches to calculate it. Both are important, and we recapitulate them after.

2.4.1 Multiplicative formula

$\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X}$ Let us denote $\boldsymbol{A} = (\boldsymbol{X}, \boldsymbol{y})$, a matrix $\boldsymbol{X}$ expanded by one column of corresponding dependent variables and and $\tilde{\boldsymbol{Z}} = \boldsymbol{A}^T\boldsymbol{A}$. Then

$$\tilde{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{X} & \boldsymbol{y}^T\boldsymbol{y} \end{bmatrix}. \tag{2.40}$$

Notice that $\boldsymbol{Z}$ is symmetric square matrix $\in \mathbb{R}^{p\times p}$, moreover we suppose that Z is regular. $\boldsymbol{X}^T\boldsymbol{y}$ is p dimensional column vector, $\boldsymbol{y}^T\boldsymbol{X}$ is p dimensional row vector and $\boldsymbol{y}^T\boldsymbol{y}$ is a scalar. OLS estimate $\hat{\boldsymbol{w}}$ is then given by

$$\hat{\boldsymbol{w}}^{(\boldsymbol{OLS},\boldsymbol{n})} = \boldsymbol{Z}^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.41}$$

and the RSS by

$$RSS = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{w}}. \tag{2.42}$$

Also, note that all necessary matrix multiplications are already in the matrix $\tilde{\boldsymbol{Z}}$.

Let us show that RSS can be expressed as ratio of determinants $\det(\tilde{\boldsymbol{Z}})$ and $\det(\boldsymbol{Z})$ (using determinant rule for block matrices) so that

$$
\begin{aligned}
RSS &= \frac{\det(\tilde{\boldsymbol{Z}})}{\det(\boldsymbol{Z})} \\
&= \frac{\det\begin{pmatrix} \boldsymbol{Z} & \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{X} & \boldsymbol{y}^T\boldsymbol{y} \end{pmatrix}}{\det(\boldsymbol{M})} \\
&= \frac{\det(\boldsymbol{Z})\det\left(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{Z}^{-1}\boldsymbol{y}\right)}{\det(\boldsymbol{Z})} \\
&= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{w}}.
\end{aligned}
\tag{2.43}
$$

If we assume that $RSS > 0$ then $\tilde{\boldsymbol{Z}}^{-1}$ can be expressed as

$$
\tilde{\boldsymbol{Z}}^{-1} = \begin{bmatrix} \boldsymbol{Z}^{-1} + \dfrac{\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}}{RSS} & -\dfrac{\hat{\boldsymbol{w}}^T}{RSS} \\ -\dfrac{\hat{\boldsymbol{w}}^{\boldsymbol{T}}}{RSS} & \dfrac{1}{RSS} \end{bmatrix}. \tag{2.44}
$$

For following equations it is important to notice that for any two row vectors $\boldsymbol{c_i} = (\boldsymbol{x_i}, y_i)$ and $\boldsymbol{c_j} = (\boldsymbol{x_j}, y_j)$, $\boldsymbol{c_i}, \boldsymbol{c_j} \in \mathbb{R}^{1\times p+1}$ is

$$
\boldsymbol{c_i}\tilde{\boldsymbol{Z}}^{-1}\boldsymbol{c_i^T} = \frac{(y_i - \boldsymbol{x_i}\hat{\boldsymbol{w}})^2}{RSS} + \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T} \tag{2.45}
$$

and

$$
\boldsymbol{c_j}\tilde{\boldsymbol{Z}}^{-1}\boldsymbol{c_j^T} = \frac{(y_j - \boldsymbol{x_j}\hat{\boldsymbol{w}})(y_i - \boldsymbol{x_i}\hat{\boldsymbol{w}})}{RSS} + \boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T} \tag{2.46}
$$

Including the observation

Using above let us express how the determinant $\det(\boldsymbol{Z})$ and the inverse of the $vecZ^{-1}$ changes when observation $\boldsymbol{c_i} = (\boldsymbol{x_i}, y_i)$ is added to the matrix $\boldsymbol{A}$. First let us notice that if we add this row to $\boldsymbol{A}$, then $\boldsymbol{Z}$ changes as

$$
\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} & x_{i_1} \\ x_{21} & x_{22} & \dots & x_{2n} & x_{i_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} & x_{i_p} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \\ x_{i_1} & x_{i_2} & \dots & x_{i_p} \end{bmatrix} = \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{x_i^T}\boldsymbol{x_i} = \boldsymbol{Z} + \boldsymbol{x_i^T}\boldsymbol{x_i}, \tag{2.47}
$$

so determinant with appended row changes as

$$
\det(\boldsymbol{Z} + \boldsymbol{x_i^T}\boldsymbol{x_i}) = det(\boldsymbol{Z})(1 + \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T}). \tag{2.48}
$$

Finally the inversion $\boldsymbol{Z}^{-1}$ can be obtained using Sherman-Morrison formula [19] so that

$$(\boldsymbol{Z} + \boldsymbol{x_i^T x_i})^{-1} = \boldsymbol{Z}^{-1} - \frac{\boldsymbol{Z}^{-1}\boldsymbol{x_i^T x_i Z}^{-1}}{1 + \boldsymbol{x_i Z}^{-1}\boldsymbol{x_i^T}} \tag{2.49}$$

It is now convenient to denote

$$b = \frac{-1}{(1 + \boldsymbol{x_i Z}^{-1}\boldsymbol{x_i^T})}, b \in \mathbb{R}, \tag{2.50}$$

and

$$\boldsymbol{u} = \boldsymbol{Z}^{-1}\boldsymbol{x_i^T}, \boldsymbol{u} \in \mathbb{R}^{p\times 1}. \tag{2.51}$$

Then (2.49) can be written as

$$(\boldsymbol{Z} + \boldsymbol{x_i^T x_i})^{-1} = \boldsymbol{Z}^{-1} + b\boldsymbol{u}\boldsymbol{u}^{\boldsymbol{T}}. \tag{2.52}$$

Given that last missing piece to express updated $\hat{\boldsymbol{w}}$ is appended $\boldsymbol{X}^T\boldsymbol{y}$ by one row, which can be simply expressed by same idea as (2.47) so that updated $\hat{\boldsymbol{w}}$ which we denote as $\overline{\hat{\boldsymbol{w}}}$ is

$$\overline{\hat{\boldsymbol{w}}} = (\boldsymbol{Z}^{-1} + b\boldsymbol{u}\boldsymbol{u}^{\boldsymbol{T}})(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} + y_i\boldsymbol{x_i^T}). \tag{2.53}$$

mention the mistake in original paper? w + (y - xw)bu ??

This can be simplified so that we get

$$\overline{\hat{\boldsymbol{w}}} = \hat{\boldsymbol{w}} - (y_i - \boldsymbol{x_i}\hat{\boldsymbol{w}})b\boldsymbol{u}. \tag{2.54}$$

Least but not last we want to express updated RSS which we denote as $\overline{RSS}$. This can be done easily from 2.43 and 2.48 as

$$\overline{RSS} = RSS + \frac{(y_i - \boldsymbol{x_i}\hat{\boldsymbol{w}})^{\boldsymbol{2}}}{(1 + \boldsymbol{x_i Z}^{-1}\boldsymbol{x_i^T})}. \tag{2.55}$$

It is convenient to mark

$$\gamma^+(\boldsymbol{c_i}) = \frac{(y_i - \boldsymbol{x_i}\hat{\boldsymbol{w}})^{\boldsymbol{2}}}{(1 + \boldsymbol{x_i Z}^{-1}\boldsymbol{x_i^T})}, \tag{2.56}$$

so that

$$\overline{RSS} = RSS + \gamma^+(\boldsymbol{c_i}) \tag{2.57}$$

We can see that $\gamma^+(\boldsymbol{c_i})$ measures how RSS increase after we append our dataset with observation $\boldsymbol{c_i}$, thus $\gamma^+(\boldsymbol{c_i}) \geq 0$

Excluding the observation

Because we want to express both increment and decrement change in our dataset, let us now focus how RSS, $\boldsymbol{Z^{-1}}$ and $\hat{\boldsymbol{w}}$ changes after we exclude one observation.

Consider that we already included one observation $\boldsymbol{c_i}$ in our dataset and mark $\overline{\boldsymbol{Z}} = \boldsymbol{Z} + \boldsymbol{x_i x_i^T}$. If we exclude one observation $\boldsymbol{c_j} = (\boldsymbol{x_j}, y_j) \in \mathbb{R}^{p+1\times 1}$ from already updated matrix $\boldsymbol{A}$ then the determinant $\det(\overline{\boldsymbol{Z}})$ changes as

$$\det(\overline{\boldsymbol{Z}} - \boldsymbol{x_j^T x_j}) = det(\overline{\boldsymbol{Z}})(1 - \boldsymbol{x_j}\overline{\boldsymbol{Z}}^{-1}\boldsymbol{x_j^T}) \tag{2.58}$$

and the inversion changes (again, according to Sherman-Morrison formula) as

$$(\overline{\boldsymbol{Z}} - \boldsymbol{x_j^T x_j})^{-1} = \overline{\boldsymbol{Z}}^{-1} + \frac{\overline{\boldsymbol{Z}}^{-1}\boldsymbol{x_j^T x_j}\overline{\boldsymbol{Z}}^{-1}}{1 - \boldsymbol{x_j}\overline{\boldsymbol{Z}}^{-1}\boldsymbol{x_j^T}}. \tag{2.59}$$

Once again, it is convenient to denote

$$\overline{b} = \frac{-1}{(1\boldsymbol{x_j}\overline{\boldsymbol{Z}}^{-1}\boldsymbol{x_j^T})}, \overline{b} \in \mathbb{R}, \tag{2.60}$$

and

$$\overline{\boldsymbol{u}} = \overline{\boldsymbol{Z}}^{-1}\boldsymbol{x_j^T}, \overline{\boldsymbol{u}} \in \mathbb{R}^{p\times 1}, \tag{2.61}$$

so that we can write

$$(\overline{\boldsymbol{Z}} + \boldsymbol{x_j^T x_j})^{-1} = \boldsymbol{Z}^{-1} - \overline{b}\overline{\boldsymbol{u}}\overline{\boldsymbol{u}}^{\boldsymbol{T}}. \tag{2.62}$$

Using the same approach as before we can denote express down-dated estimate which we denote as $\overline{\overline{\hat{\boldsymbol{w}}}}$

$$\overline{\overline{\hat{\boldsymbol{w}}}} = +\overline{\hat{\boldsymbol{w}}}(y_j - \boldsymbol{x_j}\overline{\hat{\boldsymbol{w}}})\overline{b}\overline{\boldsymbol{u}}. \tag{2.63}$$

Finally lets also express updated $\overline{RSS}$ which we denote as $\overline{\overline{RSS}}$. This can be done easily from 2.43 and 2.48 and (2.62) as

$$\overline{\overline{RSS}} = \overline{RSS} - \frac{(y_j - \boldsymbol{x_j}\overline{\hat{\boldsymbol{w}}})^{\boldsymbol{2}}}{(1 - \boldsymbol{x_j}\overline{\boldsymbol{Z}}^{-1}\boldsymbol{x_j^T})}. \tag{2.64}$$

Let us mark

$$\gamma^{-}(\boldsymbol{c_j}) = \frac{(y_j - \boldsymbol{x_j}\overline{\hat{\boldsymbol{w}}})^{\boldsymbol{2}}}{(1 - \boldsymbol{x_j}\overline{\boldsymbol{Z}}^{-1}\boldsymbol{x_j^T})}, \tag{2.65}$$

so that

$$\overline{\overline{RSS}} = \overline{RSS} - \gamma^{-}(\boldsymbol{c_j}). \tag{2.66}$$

Swapping two observations

Let us now express the equation for including and excluding observation at once. First, notice that from (2.62) we can express

$$\boldsymbol{x_j}\overline{\boldsymbol{Z}}^{-1}\boldsymbol{x_j^T}) = \boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}) - \frac{(\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T})^2}{1+\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T})}, \tag{2.67}$$

so that

$$\det(\boldsymbol{Z}+\boldsymbol{x_i^T}\boldsymbol{x_i}-\boldsymbol{x_j^T}\boldsymbol{x_j}) = \tag{2.68}$$
$$\det(\boldsymbol{Z})(1+\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T}-\boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}+(\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T})^2-\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T}\boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T})$$

Finally we can express the $\overline{\overline{RSS}}$ as

$$\overline{\overline{RSS}} = RSS\rho(\boldsymbol{c_i},\boldsymbol{c_j}), \tag{2.69}$$

where

$$\rho(\boldsymbol{c_i},\boldsymbol{c_j}) = \frac{(1+\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T}+\frac{e_j^2}{RSS})(1-\boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}-\frac{e_i^2}{RSS})+(\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}+\frac{e_ie_j}{RSS})^2}{1+\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T}-\boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}+(\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T})^2-\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T}\boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}}, \tag{2.70}$$

where $e_i = y_i - x_i\hat{\boldsymbol{w}}$ and $e_j = y_j - x_j\hat{\boldsymbol{w}}$.

We can see that this formula is similar to the (2.39) but here the $\rho(\boldsymbol{c_i},\boldsymbol{c_j})$ represents multiplicative increment. Moreover $0 < \rho(\boldsymbol{c_i},\boldsymbol{c_j})$ and if $0 < \rho(\boldsymbol{c_i},\boldsymbol{c_j}) < 1$ then the swap leads to improvement in terms of feasible solution.

We are now able to modify the FSA so that we do not need to recompute $\hat{\boldsymbol{w}}$ and inversion $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ but we can update it. Authors call this algorithm optimal exchange algorithm (OEA)

2.4.2 The OEA and its properties

We can apply theory from the previous section to the FSA. Sets O_m and Z_m (see section 2.3) contains indexes of observations to exclude and include respectively. The matrix $\boldsymbol{Z_M}$ can be used instead of the matrix $\boldsymbol{Z}$ and $\boldsymbol{X_M}$ with $\boldsymbol{y_M}$ to calculate $\hat{\boldsymbol{w}}$.

In terms of the FSA 6 we first need to change the OSAA. There are only minor tweaks. First, we need to pass one more argument, and that is RSS. Second, We want to find minimal $\rho(\boldsymbol{c_i},\boldsymbol{c_j})$ calculated by (2.70) so that $0 < \rho(\boldsymbol{c_i},\boldsymbol{c_j}) < 1$. Other parts of the algorithm remains the same. Time complexity thus remains the same.

This algorithm produces $\rho(\boldsymbol{c_i}$, and indexes $i_{swap} \in O_m$ and $j_{swap} \in Z_m$ of observations we want to swap. Given that, we can update the RSS by (2.69), $\boldsymbol{Z_M^{-1}}$ by (2.52) and (2.62) and finally update $\hat{\boldsymbol{w}}$ by (2.54) and (2.63).

Let us now talk about the time complexity of such a solution. As we said, the time complexity of modifies OSAA is $\mathcal{O}(n^2p^2)$. Thus there is no asymptotic improvement. On the other hand, there are some significant constants outside the OSAA so let us look at how they improve.

When an improvement is found, thus when $0 < \rho(\boldsymbol{c_i}, \boldsymbol{c_j}) < 1$, then we update RSS, which is constant. Next we update inversion by (2.52) which is $\mathcal{O}(4p^2)$ and (2.62) which is also $\mathcal{O}(4p^2)$. Finally we need to update $\hat{\boldsymbol{w}}$ by (2.54) and (2.63) which are both $\mathcal{O}(p^2 + p)$.

Observation 27. Time complexity of updating all the quantities is $\mathcal{O}(8p^2 + 2p^2 + p) \sim \mathcal{O}(p^2)$. That is quite an improvement if we compare it to time complexity $\mathcal{O}(p^2n)$ of updating those quantities in the FSA (see observation 26).

Note that right now it actually does not matter if we use additive formula (2.39) with the stopping criterion $\Delta S_{i,j}^{(\boldsymbol{m})} \geq 0$—thus unmodified OSAA—or multiplicative formula (2.70) with the stopping criterion $\rho(\boldsymbol{c_i}, \boldsymbol{c_j}) \geq 1$. Both results we can update RSS and other quantities can be used to calculate both formulas.

However, the advantage of the multiplicative formula 2.70 is that we can use the following bounding condition to improve the performance of the modified OSAA.

Bounding condition improvement

The $\rho(\boldsymbol{c_i}, \boldsymbol{c_j})$ is expressed as a ratio. We can see that in the numerator we have

$$(1 + \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T} + \frac{e_j^2}{RSS})(1 - \boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T} - \frac{e_i^2}{RSS}) + (\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T} + \frac{e_ie_j}{RSS})^2, \quad (2.71)$$

and because $\frac{e_ie_j}{RSS})^2 \geq 0$ then whole numerator is greater or equal to

$$(1 + \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T} + \frac{e_j^2}{RSS})(1 - \boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T} - \frac{e_i^2}{RSS}). \quad (2.72)$$

On the other hand, we can see that denominator is

$$1 + \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T} - \boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T} + (\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T})^2 - \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T}\boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}, \quad (2.73)$$

and because $(\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T})^2$ and $\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T}\boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}$ are actually inner products of $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ ($\boldsymbol{Z}^{-1}$ is positive definite) thus

$$(\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T})^2 \leq \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T}\boldsymbol{x_j} \quad (2.74)$$

using the Cauchy-Schwarz inequality. This means that denominator is less or equal to

$$1 + \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T} - \boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}. \quad (2.75)$$

Given that we can denote $\rho_b(\boldsymbol{c_i}, \boldsymbol{c_j})$ as

$$\rho_b(\boldsymbol{c_i}, \boldsymbol{c_j}) = \frac{(1 + \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T} + \frac{e_j^2}{RSS})(1 - \boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T} - \frac{e_i^2}{RSS})}{1 + \boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_i^T} - \boldsymbol{x_j}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}} \quad \leq \rho(\boldsymbol{c_i}, \boldsymbol{c_j}). \tag{2.76}$$

The actual speed improvement is then given by that we do not need to compute $\boldsymbol{x_i}\boldsymbol{Z}^{-1}\boldsymbol{x_j^T}$ in each of $h(n-h)$ pairs swap comparison. As we know this quantity cannot be computed outside the loop; thus it is the reason for such a high time complexity. The modified OSAA can be further modified as follows.

First we set $\rho_{min} := 1$. Then for each pair we only compute $\rho_b(\boldsymbol{c_i}, \boldsymbol{c_j})$ and if it is greater of equal to ρ_{min} we can continue to next pair without computing $\rho(\boldsymbol{c_i}, \boldsymbol{c_j})$. It is very useful because all quantities necessary for calculating $\rho_b(\boldsymbol{c_i}, \boldsymbol{c_j})$ can be precalculated outside of the loop.

If $\rho_b(\boldsymbol{c_i}, \boldsymbol{c_j})$ is less that ρ_{min} we actually compute $\rho(\boldsymbol{c_i}, \boldsymbol{c_j})$ and set $\rho_{min} := \rho(\boldsymbol{c_i}, \boldsymbol{c_j})$.

That means in the double loop which time complexity is $\backslash^{\in}$ we do not always need to calculate $\rho(\boldsymbol{c_i}, \boldsymbol{c_j})$ which time complexity is $\backslash^{\in}$. This does not improve the asymptotic time complexity, but (as can be seen in the table) it can improve the speed of the algorithm.

create experiment

Finally, let us note that [3] calls this algorithm with bounding condition as *modified optimal exchange algorithm* (MOEA).

2.4.3 Minimum-maximum exchange algorithm

The minimum-maximum exchange algorithm (MMEA) is very similar to the FSA respectively to its modified version the OEA. The main difference is the greediness of this algorithm. What we mean by that is that this algorithm does not find the optimal swap, but rather first find the $\boldsymbol{c_j}$ which increase the RSS by the minimum value and include this observation. Next, it finds $\boldsymbol{c_j}$ such that in decrease RSS by maximum and exclude this observation.

The minimum increase can be found by calculating $\gamma^+(\boldsymbol{c_i})$ by 2.56 for each trimmed observation in Z_m.

Next, this observation is included, so we get $h+1$ untrimmed observations. We update $\boldsymbol{Z_M^{-1}}$ to $\overline{\boldsymbol{Z_M}}^{-1}$ by (2.52) and and $\hat{\boldsymbol{w}}$ to $\overline{\hat{\boldsymbol{w}}}$ by (2.54).

Then we can find maximum $\gamma^-(\boldsymbol{c_j})$ by (2.65) among $O_m^{(i)}$ where $m^{(i)}$ denotes included observation c_i.

Next, we can update $\overline{\boldsymbol{Z_M}}^{-1}$ to the $\overline{\overline{\boldsymbol{Z_M}}}^{-1}$ by (2.62) and $\overline{\hat{\boldsymbol{w}}}$ to $\overline{\overline{\hat{\boldsymbol{w}}}}$ by (2.63).

Finally, we can update RSS to $\overline{\overline{RSS}}$ by 2.57 and 2.66. We can repeat this process until $\gamma^-(\boldsymbol{c_j}) > \gamma^+(\boldsymbol{c_i})$.

Observation 28. One step of the algorithm MMEA has time complexity $O(p^2n)$. So the whole algorithm has time complexity $\mathcal{O}(tlp^2n)$ where t is number of the starts and l is maximum number of iterations of each start.

Proof. Because we do not iterate through all pairs but only over $n-h$ followed by $h+1$ subsets the loops takes only $\mathcal{O}(n)$ time. In the loops, we are computing $\gamma^-(\boldsymbol{c_j})$ and $\gamma^+(\boldsymbol{c_i})$ both takes $\mathcal{O}(p^2)$ time. Outside the loops, all the each of the quantities we are updating take $\mathcal{O}(p^2)$ time. That means the one loop of the whole algorithm last $\mathcal{O}(p^2n)$. As in the case of the FSA, if we introduce parameters t and l, then the time complexity of the algorithm is $\mathcal{O}(tlp^2n)$ □

2.4.4 Different method of computation

In the last section we introduced a way of calculating the OEA so that we update $\hat{\boldsymbol{w}}$ and inversion $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. This, however, requires to compute inversion at the start of the algorithm (this is also the case for the FSA, MOEA and MMEA) As we know, calculating inversion is not practical due to low numerical stability. In practice, we usually use QR decomposition. In this section, we describe how we can modify OEA to use the QR decomposition (the same idea can also be applied to the FSA, MOEA and the MMEA).

Let us start by describing how we can update the QR factorization, which is a critical part of this modified computation. Assume that we have QR decomposition of $\boldsymbol{X}$, and we need to exchange ith observation from O_m with jth observation from Z_m. We can simulate this by inserting jth row and consequently extracting ith row from the QR decomposition.

QR insert

First, let us discuss how to update QR decomposition when a row $\boldsymbol{x_j}$ is inserted. If we add a row to $\boldsymbol{A}$ as the last row, our decomposition looks like

$$\overline{\boldsymbol{R}} = \overline{\boldsymbol{Q}}\boldsymbol{A}^{(+)} = \begin{bmatrix} \times & \cdots & \cdots & \cdots & \times \\ 0 & \times & \cdots & \cdots & \times \\ \vdots & 0 & \ddots & \cdots & \vdots \\ \vdots & \vdots & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & 0 & \times \\ \vdots & \vdots & \vdots & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \times & \cdots & \cdots & \cdots & \times \end{bmatrix} \tag{2.77}$$

Where $\overline{\boldsymbol{R}}$ denotes matrix $\boldsymbol{R}$ which is now not upper triangular and needs to be updated and $\overline{\boldsymbol{Q}} \in \mathbb{R}^{p+1\times p+1}$ denotes matrix created from $\boldsymbol{Q}$ by adding one row and one column with zeros. To preserve zeroes on the diagonal we in addition

put $\overline{\boldsymbol{Q}}_{n+1,n+1} = 1$. To update $\overline{\boldsymbol{R}}$ to upper triangular matrix $\boldsymbol{R}^{(+)}$, we need to create additional orthogonal Givens matrices $\boldsymbol{Q_{e+1}} \dots \boldsymbol{Q_{e+p}}$. Then updated upper triangular matrix is $\boldsymbol{R}^{(+)} = \boldsymbol{Q_{e+p}} \dots \boldsymbol{Q_{e+2}} \boldsymbol{Q_{e+1}} \boldsymbol{R}$.

Updated matrix $\overline{\boldsymbol{Q}}$ which we denote as $\boldsymbol{Q}^{(+)}$ is updated in the same manner, thus $\boldsymbol{Q}^{(+)} = \boldsymbol{Q} \boldsymbol{Q_e^T} \boldsymbol{Q_{e+1}^T} \dots \boldsymbol{Q_{e+p}^T}$. Note that if we do not want to have inserted row $\boldsymbol{x_j}$ as the last but as (say kth) row, then in terms of QR decomposition, we only need to move (last) $\boldsymbol{x_j}$ row to the kth position. In other words, we create a permutation matrix $\boldsymbol{P}$ so that

$$\boldsymbol{P}\boldsymbol{Q}^{(+)} = \begin{bmatrix} \boldsymbol{A}(1:k-1, 1:p) \\ \boldsymbol{x_j} \\ \boldsymbol{A}(1:k+1, 1:p), \end{bmatrix} \tag{2.78}$$

where $\boldsymbol{A}(a:b, 1:p)$ denotes rows of matrix $\boldsymbol{A}$ from a to b. We can describe the whole process by the following pseudocode.

Algorithm 7: QR insert

Input: $\boldsymbol{Q} \in \mathbb{R}^{n \times n}, \boldsymbol{R} \in \mathbb{R}^{n \times p}, \boldsymbol{x_j} \in \mathbb{R}^p, k$
Output: $\boldsymbol{Q}^{(+)} \in \mathbb{R}^{n+1 \times n+1}, \boldsymbol{R}^{(+)} \in \mathbb{R}^{n+1 \times p}$

$\boldsymbol{R}^{(+)} \leftarrow \boldsymbol{R}$ *with appended last row by* $\boldsymbol{x_j}$;
$\boldsymbol{Q}^{(+)} \leftarrow \boldsymbol{Q}$ *with appended last row and last column by zeors*;
$\boldsymbol{Q}^{(+)}{}_{n+1,n+1} \leftarrow 1$; // i.e. $Q^{(+)}$ now has 1 on the diagonal
for $i \leftarrow n$ ***to*** p **do**
  $\boldsymbol{Q}(i, n+1) \leftarrow$ *create Givens matrix* $\boldsymbol{Q}(i, n+1) \in \mathbb{R}^{n+1 \times n+1}$;
  $\boldsymbol{R}^{(+)} \leftarrow \boldsymbol{Q}(i, n+1) \boldsymbol{R}^{(+)}$;
  $\boldsymbol{Q}^{(+)} \leftarrow \boldsymbol{Q}^{(+)} \boldsymbol{Q}^T(i, n+1)$;
end
for $i \leftarrow n+1$ ***to*** k **do**
  $\boldsymbol{Q}^{(+)} \leftarrow \boldsymbol{Q}^{(+)}$ *where we swap the* i *row with the* $i-1$ *row*
end
return $\boldsymbol{Q}^{(+)}$, $\boldsymbol{R}^{(+)}$;

Observation 29. Computing $\boldsymbol{R}^{(+)}$ is $\mathcal{O}(p^2)$ and computing $\boldsymbol{Q}^{(+)}$ is $\mathcal{O}(np)$.

Proof. We have loop over p $\boldsymbol{Q}(i, j)$ Givens matrices. We do not need to create those matrices, because by multiplying $\boldsymbol{R}^{(+)}$ or $\boldsymbol{Q}^{(+)}$ with the matrix $\boldsymbol{Q}(i, j)$ only one row of is affected. That means we can simulate this matrix-matrix multiplication only by iterating over those matrices and multiplying elements with $\cos(\varphi)$ and $\sin(\varphi)$ adequately. This means in case of $\boldsymbol{R}^{(+)}$ we are iterating p times over nonzero rows of $\boldsymbol{R}^{(+)}$ thus p rows. That gives us time complexity of $\mathcal{O}(p^2)$. In the case of $\boldsymbol{Q}^{(+)}$ we are iterating p times over columns of $\boldsymbol{Q}^{(+)}$ and that gives us time complexity of $\mathcal{O}(np)$. □

Note 30. This process can also be done in case we are using an economic version of matrices $\boldsymbol{R}$ and $\boldsymbol{Q}$ thus matrices $\boldsymbol{R_1}$ and $\boldsymbol{Q_1}$ (see (2.11)). In such a case $\boldsymbol{Q_1} \in \mathbb{R}^{p \times p}$ thus then updating $\boldsymbol{Q_1}$ to $\boldsymbol{Q_1}^{(+)}$ is only $\mathcal{O}(p^2)$.

Note 31. The matrix $\overline{\boldsymbol{R}}$ can be updated to $\boldsymbol{R}^{(+)}$ without the presence of matrix $\boldsymbol{Q}$. We use this observation in the algorithm described in 2.6.

QR delete

When we extract the row $\boldsymbol{x_i}$ from matrix $\boldsymbol{A}$ we can use following trick [15]. First, we move such row as the first row of the matrix $\boldsymbol{A}$ so ve create permutation matrix $\boldsymbol{P}$ so that

$$\boldsymbol{PA} = \begin{bmatrix} \boldsymbol{x_i} \\ \boldsymbol{A}(1:i-1,1:p) \\ \boldsymbol{A}(1:i+1,1:p) \end{bmatrix} = \begin{bmatrix} \boldsymbol{x_i} \\ \boldsymbol{A}^{(-)} \end{bmatrix} = \boldsymbol{PQR} \tag{2.79}$$

where $\boldsymbol{A}(a:b,1:p)$ means rows of matrix $\boldsymbol{A}$ from a to b. We can see that we only need to introduce zeros in the first row $\boldsymbol{q_1}$ (except q_{11}) of the matrix $\boldsymbol{Q}$. We can do this by $n-1$ matrices $\boldsymbol{Q}(i,j) \in \mathbb{R}^{n \times n}$ so that

$$\boldsymbol{Q}(1,2)\dots\boldsymbol{Q}(n-1,n)\boldsymbol{q_1^T} = \begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix} \tag{2.80}$$

To propagate the change into $\boldsymbol{R}$ we then consequently need to update $\boldsymbol{R}$ so that

$$\boldsymbol{Q}(1,2)\dots\boldsymbol{Q}(n-1,n)\boldsymbol{R} = \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{R}^{(-)} \end{bmatrix}. \tag{2.81}$$

The result is then

$$\boldsymbol{PA} = (\boldsymbol{PQQ^T}(n-1,n)\dots\boldsymbol{Q^T}(1,2))(\boldsymbol{Q}(1,2)\dots\boldsymbol{Q}(n-1,n)\boldsymbol{R}) = \begin{bmatrix} 1 & 0 \\ 0 & \boldsymbol{Q}^{(-)} \end{bmatrix} \begin{bmatrix} \times \\ \boldsymbol{R}^{-} \end{bmatrix}, \tag{2.82}$$

so that

$$\boldsymbol{A}^{(-)} = \boldsymbol{Q}^{(-)}\boldsymbol{R}^{(-)} \tag{2.83}$$

Pseudocode is very similar to the QR Insert algorithm 7 so it is not necessary to provide it here.

Observation 32. Time complexity of QR delete is $\mathcal{O}(n^2)$.

Proof. We need to create $n-1$ Givens matrices, and with each of this we need to multiply $\boldsymbol{Q}^{(-)}$ and $\boldsymbol{R}^{(-)}$. As we stated in 12 this can be node in $\mathcal{O}(n)$ for each matrix. So together we get $\mathcal{O}(n^2)$. □

Note 33. In case of QR delete, it is not possible to use the economic version of matrices.

Calculation of OEA using QR decomposition

Given all the required theory above, let us describe the computation. Let us start with (2.70). Here we need inversion to calculate $\boldsymbol{x_i Z^{-1} x_i^T}$, $\boldsymbol{x_j Z^{-1} x_j^T}$ and $\boldsymbol{x_i Z^{-1} x_j^T}$. But this can also be done without inversion, only using the decomposition. We can write this equation

$$\boldsymbol{x_i Z^{-1} x_i^T} = \boldsymbol{v}^T\boldsymbol{v} \iff \boldsymbol{x_i}(\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{x_i^T} = \boldsymbol{v}^T\boldsymbol{v} \tag{2.84}$$

where $\boldsymbol{v}$ can be obtained by solving lower triangular system

$$\boldsymbol{R}^T\boldsymbol{v} = \boldsymbol{x_i}^T. \tag{2.85}$$

The same can be done with $\boldsymbol{x_j Z^{-1} x_j^T}$. Least but not last we need to solve $\boldsymbol{x_i Z^{-1} x_j^T}$. We can write this

$$\boldsymbol{x_i Z^{-1} x_j^T} = \boldsymbol{x_j u} \iff \boldsymbol{x_i}(\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{x_j^T} = \boldsymbol{x_j}^T\boldsymbol{u}, \tag{2.86}$$

where column vector $\boldsymbol{u}$ is defined by (2.51). We can see that

$$\boldsymbol{u} = \boldsymbol{Z^{-1} x_i^T} \iff (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{x_i^T} = \boldsymbol{u} \tag{2.87}$$

so that $\boldsymbol{u}$ can be obtained by solving the upper triangular system

$$\boldsymbol{Ru} = \boldsymbol{v}, \tag{2.88}$$

where $\boldsymbol{v}$ is solution of (2.85). Other quantities of (2.70) does not require $\boldsymbol{Z^{-1}}$.

Observation 34. We can see that time complexity of all these calculations is the same as in case of calculating with inversion thus $\mathcal{O}(p^2)$.

Proof. Indeed, in the case of the inversion, we are multiplying quantities such as $\boldsymbol{x_i Z^{-1} x_i^T}$ which can be done in $\mathcal{O}(p^2)$. In the case of decomposition, we solve this problem by solving the triangular system of p equations. This can also be done in $\mathcal{O}(p^2)$. □

When optimal exchange $\boldsymbol{c_i}, \boldsymbol{c_j}$ is found then we need to update RSS which can be done same way by (2.69). Updating $\hat{\boldsymbol{w}}$ to $\overline{\hat{\boldsymbol{w}}}$ by (2.54) requires $\boldsymbol{u}$ which in this case we calculate by (2.88) and b which requires $\boldsymbol{x_j Z^{-1} x_j^T}$. This can be done in the same manner as (2.84). Analogous operations can be used to update $\overline{\hat{\boldsymbol{w}}}$ to $\overline{\overline{\hat{\boldsymbol{w}}}}$ by means of (2.63). Only thing wee need to realize that we can express (2.67) as $\boldsymbol{x_j}\overline{\boldsymbol{Z}}^{-1}\boldsymbol{x_j^T} = \boldsymbol{x_i Z^{-1} x_i^T} + b(\boldsymbol{u x_j}^2)$.

On the other hand, we cannot use equation (2.52) and (2.62) for updating the inversion, because we do not have one. So we need to update our QR decomposition somehow. We describe algorithms for updating the QR decomposition in the section 2.4.4. First, we can use the QR Insert algorithm to add row $\boldsymbol{x_j}$ to the decomposition and consequently QR delete to extract $\boldsymbol{x_j}$ from the decomposition.

Observation 35. This approach is slower than updating inversion directly. On the other hand, this solution is numerically stable. Updating inversion can be done in $\sqrt{}^{\in}$ (see observation 27) while QR insert has $\mathcal{O}(np)$ (see observation 29) and time complexity and QR delete even $\mathcal{O}(n^2)$ (see observation 32).

Note 36. Because time complexity of the QR delete is $\mathcal{O}(n^2)$, it is up to consideration if instead of recycling QR decomposition is not worth it to re-calculate it from scratch which takes $\mathcal{O}(p^2 n)$ (see (2.20)).

Finally let us talk about matrix $\tilde{\boldsymbol{Z}}$ (2.40) which we used for derivation of our equations. If we use 2.19 observation then we realize that if we make QR factorization of $\boldsymbol{A} = (\boldsymbol{X}, \boldsymbol{y})$ then

$$\tilde{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{X} & \boldsymbol{y}^T\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}^{\boldsymbol{T}} & 0 \\ \boldsymbol{\phi}^{\boldsymbol{T}} & r \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{\phi} \\ 0 & r \end{bmatrix} \tag{2.89}$$

where

$$\tilde{\boldsymbol{R}} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{\phi} \\ 0 & r \end{bmatrix} = \tilde{\boldsymbol{Q}^{\boldsymbol{T}}}\boldsymbol{A} \tag{2.90}$$

is matrix from QR factorization of $\boldsymbol{A}$. Next, we can realize that $\boldsymbol{R} \in \mathbb{R}^{p \times p}$ is matrix $\boldsymbol{R}$ which we can obtain by QR factorization of $\boldsymbol{X}$. Moreover $\boldsymbol{\phi} \in \mathbb{R}^{p \times 1}$ is column vector which is actually equal to

$$\boldsymbol{\phi} = \boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{y} \tag{2.91}$$

where $\boldsymbol{Q}$ is matrix $\boldsymbol{Q}$ from QR factorization of $\boldsymbol{R}$. Due to this fact $\hat{\boldsymbol{w}}$ is solution of upper triangular system

$$\boldsymbol{R}\hat{\boldsymbol{w}} = \boldsymbol{\phi} \tag{2.92}$$

Finally $r \in \mathbb{R}$ is scalar such that

$$r^2 = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{w}} = RSS \tag{2.93}$$

Remark 37. We can see that all the quantities we use in the algorithm can be obtained from matrix $\tilde{\boldsymbol{R}}$.

Now we have described both versions of calculation the algorithm OEA. As we already stated, it easy to use this approach on algorithms FSA, MOEA, and MMEA. Using inversion $\boldsymbol{Z}^{-1}$ is slightly faster, but numerically less stable, second, using QR decomposition is more numerically stable. Asymptotically both approaches provide the same performance. Finally, we have also shown that using matrix $\tilde{\boldsymbol{R}}$ is useful because it contains all necessary quantities for the algorithms.

2.5 Combined algorithm

In this section, we shortly describe how we can utilize lemma 12; thus that strong necessary condition is not satisfied unless the weak necessary condition is satisfied.

Algorithm FAST-LTS outputs h element subset satisfying the weak necessary condition. One step of this algorithm has time complexity $\mathcal{O}(p^2n)$ (see observation 21). The iteration of this step is quite low and moreover usually limited by parameter.

On the other hand, algorithms for finding strong necessary condition FSA, OEA and MOEA have time complexity of one step $\mathcal{O}(p^2n^2)$ (we now don't take into account eager version MMEA, because we have no proof that it finds h element subset satisfying strong necessary condition).

We can use this in our favor so that we can find h element subset satisfying weak necessary condition by the FAST-LTS algorithm and consequently use this h element subset as an input to some of the algorithms which find h element subset satisfying the strong necessary condition.

Let us consider the combination of the FAST-LTS and the MOEA. We have two options on how to perform this: z

1. Let the FAST-LTS converge and use this h element subset as input to the MOEA which we let converge.

2. Let the FAST-LTS converge to some h element subset. Perform one step of the MOEA and use the result as the input to the FAST-LTS. We can then iterate between steps of these two algorithms till convergence.

On the large data sets, where even one step of the MOEA is too exhausting, we can use the eager MMEA which time complexity of one step is lower than in the case of MOEA.

In chapter 3 we show the experimental results of various combinations of these algorithms.

2.6 BAB algorithm

In this section, we describe the algorithm from [3].Very similar algorithm can also be found in [20].

Unlike previous algorithms, this one is exact. As we discussed in section 1.4.1, exact algorithm calculate OLS fit on all of the $\binom{n}{h}$ h element subsets. For larger data sets this approach is computationally prohibitive.

This version of the exact algorithm is based on the branch and bound design paradigm; thus it tries to avoid exhaustive computation on all h element subsets. First, let us describe how the combination tree is built.

Let us denote subset of indexes $J_k = (j_1, j_2, \ldots, j_k) \subset 1, 2, \ldots, n$. We can then mark $\boldsymbol{X}_{\boldsymbol{J_k}}$ and $\boldsymbol{Y}_{\boldsymbol{J_k}}$ matrices created from $\boldsymbol{X}$ and $\boldsymbol{Y}$ so that all rows that

are not indexed by J_k are removed. We can see that the number of subsets is given by $\binom{n}{k}$.

Let us now consider tree such that at level k is $\binom{n}{k}$ nodes representing all J_k subsets. The depth of the tree is h, so this tree has $\binom{n}{h}$ leaves representing all h index subsets. Example of such a tree we can see at figure 2.3.

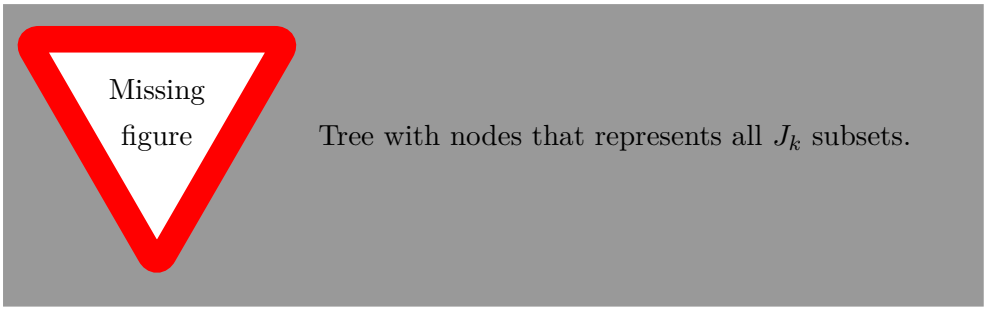


Figure 2.3: Tree consisting of all index subsets for $n = 4$ and $h = 3$

In the case of the exhaustive approach, we can then traverse the tree starting in the root using the right to left (RTL) preorder traversal, and in each node, we can calculate the OLS fit on $\boldsymbol{X}_{\boldsymbol{J_k}}$ and $\boldsymbol{Y}_{\boldsymbol{J_k}}$. If we are in the level $k = h$, we calculate the fit on the h element subset. This approach is exhaustive, so let us now describe how we can improve efficiency.

First, it is not necessary to calculate fit on kth level of the tree in a case $k < p$ because this would mean that matrix $\boldsymbol{X}_{\boldsymbol{J_k}}$ is not regular. The main improvement can because we can skip parts of the tree (perform pruning) based on the following criteria.

If at any step (at any node) of the tree traversal we calculate the $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(m_k)$ where $m_{k_i} = 1$ when $k_i \in J_k$ and $m_{k_i} = 0$ and we find out that this value is higher than minimal value found for some $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(m_h)$ thus value of OLS fit on some h element subset, then we can discard all subsets which contain J_k. That means we can trim all descendants of the node which represent J_k subset.

We can describe the algorithm as follows:

1. Set $RSS^* = \inf.$ m^*
2. For each node in RTL preorder traversal calculate $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(m_k)$
3. If $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(m_k) > RSS^*$, skip all siblings of this node in the traversal.
4. If $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(m_k) < RSS^*$ and $k = h$, then set $RSS^* = \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(m_k)$ and m^* so that $m_{k_i} = 1$ when $k_i \in J_k$ and $m_{k_i} = 0$.
5. After the traversal return m^* and $\hat{\boldsymbol{w}}$

During the traversal of the tree, we only need to know the path back to the root. Thus the whole tree does not have to be in the memory.

Remark 38. If we calculated $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(m_k)$ than we do not need to calculate $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(m_{k+1})$ from scratch. In section 2.4 we described the way how to update it when a row is added using the matrix inversion any we can apply it here. In section 2.4.4 we described the way of doing this using the matrix decomposition which is another option.

Moreover, we described that if rows are not extracted and only inserted, then matrix $\boldsymbol{Q}$ do not have to be present and only matrix $\boldsymbol{R}$ can be updated. We can use this in our favor because we can keep all the decompositions on the path from the root in the memory. If we use matrix $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{R}}$ we can (see remark 37) calculate all the quantities required for the update only using the matrix $\tilde{\boldsymbol{R}}$. This means that both versions of updating are possible and both can be done in $\mathcal{O}(p^2)$. Thus we recommend using the matrix decomposition version of the updating because it can provide higher numerical stability at the same speed.

2.6.1 Improvements

Because the tree is pruned based on the smallest value RSS^*, it would be convenient if we were able to obtain small value RSS^* early in the traversal. We can obtain this value by calculating an approximative LTS estimate. This can be done using any probabilistic algorithm described in the previous sections of this chapter. An ideal candidate may be some combined algorithm described in section 2.5. When we find such small value, it is then also convenient to permutate the observations based on its absolute residuals based on this LTS estimate in the decreasing order.

Another improvement can be made using the following sorting rule *sorting rule*: If we are at the level k (we assume that $k > p$) in the node J_k and this node has s siblings indexed as $s_1 \dots s_s$ then we can change the order of those siblings first by calculating partial increment (2.56) for each sibling and consequently ordered them in the descending order from left to right. This means we visit siblings with a lowest partial increment first. Trimming can then be done in the same manner. Moreover, if we calculate RSS in one of those siblings we and realize that $RSS > RSS^*$, then we can on the top of trimming all siblings of this sibling trim also all brothers left to this sibling. That means we can trim parent node J_k because all of his siblings have already been explored or can be trimmed. We can see such a procedure at figure 2.4.

2.7 BSA algorithm

In this section, we'll introduce another exact algorithm. It has a little different approach than previous algorithms. First of all, build a theoretical basis for

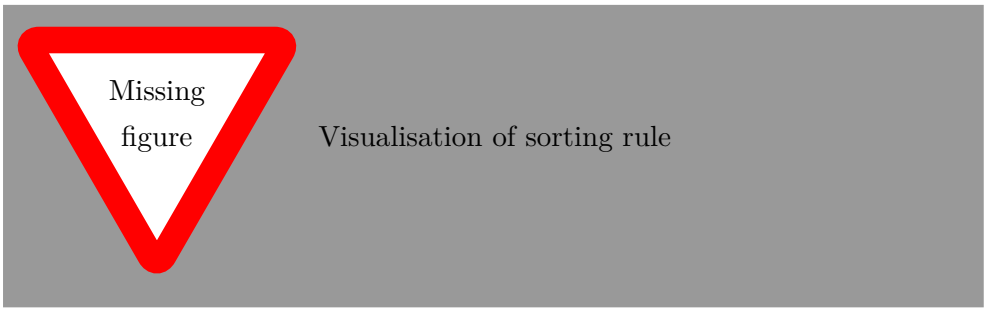


Figure 2.4: Step in the algorithms when due to the sorted siblings node j_k can be trimmed

this algorithm.

2.7.0.1 Domain of OF-LTS

We introduced two version of OF-LTS. First one is $\mathrm{OF}^{(LTS,h,n)}(\boldsymbol{w}), \boldsymbol{w} \in \mathbb{R}^p$ and the second is discrete version $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}), \boldsymbol{m} \in Q^{(n,h)}$. We also know that $\min_{\boldsymbol{w}\in\mathbb{R}^p} \mathrm{OF}^{(LTS,h,n)}(\boldsymbol{w}) = \min_{\boldsymbol{m}\in Q^{(n,h)}} \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m})$.

(see (1.27)) Let us now talk about the non-discrete version and introduce some new feature.

Definition 39. Let $Z \subset \mathbb{R}^p \times Q^{(n,h)}$ be an relation defined as

$$(\boldsymbol{w}, \boldsymbol{m}) \in Z \Leftrightarrow \sum_{i=1}^{h} r_{i:n}^2(\boldsymbol{w}) = \sum_{i=1}^{n} m_i r_i^2(\boldsymbol{w}) \tag{2.94}$$

Z is not a mapping. To show this we can take simple example so that $r_h^2 = r_{h+1}^2$. Then we have for the given $\boldsymbol{w}$ two different vectors $\boldsymbol{m}$ that are in relation with it.

For that reason, let us define $\mathcal{U} \subset \mathbb{R}^p$ as a maximal set where Z is a mapping. Next we define $\mathcal{H} = \mathbb{R}^p \setminus \mathcal{U}$ as a complement of $\mathcal{U}$.

Let's now describe some properties based on which $\boldsymbol{w}$ is in $\mathcal{U}$ or $\mathcal{H}$ set.

Theorem 40. $\boldsymbol{w} \in \mathcal{U}, \boldsymbol{w} \in \mathbb{R}^p$ if and only if $r_{(h:n)}^2(\boldsymbol{w}) < r_{(h+1:n)}^2(\boldsymbol{w})$

Proof. As shown in example above. If $r_i^2(\boldsymbol{w}) = r_{(h:n)}^2(\boldsymbol{w}) = r_{(h+1:n)}^2(\boldsymbol{w}) = r_j^2(\boldsymbol{w}), i, j \in \{1, 2, \ldots, n\}$ then $(\boldsymbol{w}, \boldsymbol{m_1}) \in Z$ and also $(\boldsymbol{w}, \boldsymbol{m_2}) \in Z$ where $\boldsymbol{m_1}$ has ones at indexes of h smallest residuals and $\boldsymbol{m_2}$ has ones at the same indexes except swap ith one with jth zero. □

Corollary 41. Complement of vectors $\boldsymbol{w}$ from theorem 40 are vectors such that

$$r_{(h:n)}^2(\boldsymbol{w}) = r_{(h+1:n)}^2(\boldsymbol{w}) \tag{2.95}$$

thus

$$\mathcal{H} = \{\boldsymbol{w} \in \mathbb{R}^p | r^2_{(h:n)}(\boldsymbol{w}) = r^2_{(h+1:n)}(\boldsymbol{w})\}. \tag{2.96}$$

That means that for each $\boldsymbol{w} \in \mathcal{H}$ there are two different $(\boldsymbol{x_i}, y_i)$ and $(\boldsymbol{x_j}, y_j)$ so that

$$(y_i - \boldsymbol{x_i w})^2 = r_i^2(\boldsymbol{w}) = r^2_{(h:n)}(\boldsymbol{w}) = r^2_{(h+1:n)}(\boldsymbol{w}) = r_j^2(\boldsymbol{w}) = (y_j - \boldsymbol{x_j w})^2. \tag{2.97}$$

We can see that

$$(y_i - \boldsymbol{x_i w})^2 = (y_j - \boldsymbol{x_j w})^2 \iff y_i \pm y_j + (\boldsymbol{x_i} \pm \boldsymbol{x_j})\boldsymbol{w} = 0 \tag{2.98}$$

Assumption 42. For $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ let as assume that for all $i, j, \in \{1, 2 \ldots, n\}$ if $i \neq j$ then $\boldsymbol{x_i} \neq \pm \boldsymbol{x_j}$ and $\|\boldsymbol{x_i}\| \neq 0$.

If assumption 42 is fulfilled then $y_i \pm y_j + (\boldsymbol{x_i} \pm \boldsymbol{x_j})\boldsymbol{w} = 0$ represents a hyperplane.

It's easy to show that set $\mathcal{U}$ is an open and Lebesgue measure of $\mathcal{H}$ is 0 [21]. Moreover $\mathcal{H}$ splits $\mathbb{R}^p$ into finite number of m open disjoint subsets of $\mathcal{U}$.

Definition 43. Let's define set of $m \in \mathbb{N}$ sets $\mathcal{U}^{(set)} = \{U_i\}_{i=1}^m$ so that all U_i are open and connected sets, U_i U_j are disjoined thus $U_i \cap U_j = \emptyset$, $\cup_{i=1}^m U_i = \mathcal{U}$ and $\cup_{i=1}^m \partial U_i = \mathcal{H}$.

We define neighbor sets $U_i, U_j, i \neq j$ so that $U_i \cap U_j \neq \emptyset$.

We also define $M^{(min)}$ set of m vectors $\boldsymbol{m} \in Q^{(n,h)}$ so that

$$M^{(min)} = \{\boldsymbol{m_1}, \ldots, \boldsymbol{m_m} | \boldsymbol{m_i} = Z(\boldsymbol{w})\boldsymbol{w} \in U_i\}.$$

where $Z(\boldsymbol{w})$ represent unique vector $\boldsymbol{m_i}$. That means $Z(\boldsymbol{w}) = Z(\boldsymbol{w'})$ for all $w, w' \in U_i$.

For a better understanding definition above see

p=1 n=5 h=3. visualise all U_i and elements of $\mathcal{H}$

We say that matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ has *h-full* if for all $m \in Q^{(n,h)}$ matrix $\boldsymbol{MX}$ has rank p.

Theorem 44. If the matrix $\boldsymbol{X}$ has h-full rank then for every local minima at point satisfying weak-necessary condition $\boldsymbol{w_0}$ of the OF-LTS function, then there exists a vector $\boldsymbol{m} \in Q^{(n,h)}$ such that $(\boldsymbol{w_0}, \boldsymbol{m}) \in Z$.

Proof can be found in [21, Theorem 7].

2.7.1 One-dimensional version of the algorithm

Based on the theorem 44 we can see that

$$\min_{\boldsymbol{m}\in Q^{(n,h)}} \mathrm{OF}^{(OLS,\boldsymbol{MX},\boldsymbol{MY})}(\boldsymbol{w}) = \min_{m\in M^{(min)}} \mathrm{OF}^{(OLS,\boldsymbol{m^iX},\boldsymbol{m^iY})}(\boldsymbol{w}) \tag{2.99}$$

where $\boldsymbol{M} = diag(\boldsymbol{m})$, $\boldsymbol{m^i} = diag(m)$. Set $M^{(min)}$ is very useful because we can iterate only through this set and not through whole $Q^{(n,h)}$.

Then minimizing objective function $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m})$ from 1.31 can be reformulated as

$$\min_{\boldsymbol{m}\in Q^{(n,h)}} \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m}) = \min_{m\in M^{(min)}} \mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(m) \tag{2.100}$$

That means if we can find all $m \in M^{(min)}$ then minimizing $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}$ would be easy. So how we can find the elements of $M^{(min)}$ set?

For now, let us assume that we know elements of $\mathcal{H}$.

Because for each $m \in M^{(min)}$ exists at least one $\boldsymbol{w} \in \mathcal{H}$ such that $(\boldsymbol{w}, \boldsymbol{m}) \in Z$ then for given $\boldsymbol{w} \in \mathcal{H}$ we can find all $\boldsymbol{m}$ by algorithm described by following pseudocode:

Algorithm 8: Find all $\boldsymbol{m}$

Input: $\boldsymbol{Q} \in \mathbb{R}^{n\times n}, \boldsymbol{R} \in \mathbb{R}^{n\times p}, \boldsymbol{x_j} \in \mathbb{R}^p, k$

Output: $\boldsymbol{Q^{(+)}} \in \mathbb{R}^{n+1\times n+1}, \boldsymbol{R^{(+)}} \in \mathbb{R}^{n+1\times p}$

write the pseudocode

1 **return** $\boldsymbol{Q^{(+)}}$, $\boldsymbol{R^{(+)}}$;

Finally we need to find all elements of $\mathcal{H}$. As we know for all elements $\boldsymbol{w} \in \mathcal{H}$ $r^2_{(h:n)}(\boldsymbol{w}) = r^2_{(h+1:n)}(\boldsymbol{w})$. So if we find all $\boldsymbol{w}$ that solve this, then we found $\mathcal{H}$. The problem is that residuals are sorted. To overcome this issue by finding some candidates which would possible be equal $(h:n)$th and $(h+1:n)$th squared residuals. In [21] we are provided necessary condition for such a candidates.

Definition 45. If $\boldsymbol{w} \in \mathcal{H}$ then there exists $i, j, i \neq j$ so that $r_i^2(\boldsymbol{w}) = r_j^2(\boldsymbol{w})$. We denote set H containing $\boldsymbol{w}$ satisfying this necessary condition, thus

$$H = \{\boldsymbol{w} \in \mathbb{R}^p | r_i^2(\boldsymbol{w}) = r_j^2(\boldsymbol{w}), i \neq j\}. \tag{2.101}$$

Note that $\binom{|H|\leq 2}{np+1}$. So in the case of the one dimensional case $\binom{|H|\leq 2}{n2=n(n-1)}$.

Finding all elements of H is then equal for for solving $\binom{}{np+1}$ equations. Moreover, because equations are quadratic, we can have up to two solutions for each equation.

So idea of the whole algorithm is to find all elements of H (by solving quadratic equations), consequently finding which of them are part of $\mathcal{H}$ (by ordering squared residuals and checking if $r^2_{(h:n)}(\boldsymbol{w}) = r^2_{(h+1:n)}(\boldsymbol{w})$ and finally for each $\boldsymbol{w} \in \mathcal{H}$ find $\boldsymbol{m}$ that are in relation. All of those $\boldsymbol{m}$ vectors form $M^{(min)}$ set. Finally we can use $\boldsymbol{m} \in M^{(min)}$ for minimizing $\mathrm{OF}_{\mathrm{D}}^{\mathrm{LTS}}(\boldsymbol{m})$ by the terms of (2.100). The whole algorithm, which authors call Border Scanning Algorithm (BSA) is described by the following pseudocode:

write the pseudocode

Algorithm 9: BSA

Input: $\boldsymbol{Q} \in \mathbb{R}^{n \times n}, \boldsymbol{R} \in \mathbb{R}^{n \times p}, \boldsymbol{x_j} \in \mathbb{R}^p, k$
Output: $\boldsymbol{Q}^{(+)} \in \mathbb{R}^{n+1 \times n+1}, \boldsymbol{R}^{(+)} \in \mathbb{R}^{n+1 \times p}$
1 **return** $\boldsymbol{Q}^{(+)}$, $\boldsymbol{R}^{(+)}$;

2.7.2 Multidimensional BSA

If we would like to scale the algorithm into the higher dimension, we have to face one major complication. The algorithm finds all elements of H and $\mathcal{H}$. This is possible in one dimension, but because $\mathcal{H}$ forms a hyperplane, then for higher dimensions is infinite.

Authors then propose to find finite set $mathcal H_p$ such that

$$\forall \boldsymbol{m} \in M^{(min)} \exists \boldsymbol{w} \in \mathcal{H}_p : (\boldsymbol{w}, \boldsymbol{m}) \in Z \tag{2.102}$$

and also finite set H_p, so that $mathcal H_p \subset H_p \subset H$. We can see that for $p = 1$ we need only one equation so the solution has dimension 0. If we would like to have some solution with dimension equal to 0 for $p > 1$ we need system of p equations. Moreover this system of equations have to be independent. So we can consider system of p equations of type $r_i^2(\boldsymbol{w}) = r_j^2(\boldsymbol{w})$

$$\begin{aligned} r_{i_1}^2(\boldsymbol{w}) &= r_{i_2}^2(\boldsymbol{w}) \\ r_{i_2}^2(\boldsymbol{w}) &= r_{i_3}^2(\boldsymbol{w}) \\ \vdots \qquad & \qquad \vdots \\ r_{i_p}^2(\boldsymbol{w}) &= r_{i_{p+1}}^2(\boldsymbol{w}) \end{aligned} \tag{2.103}$$

where $i_1, i_2 \dots i_{p+1}$ is $(p+1)$-element subset of $1, 2, \dots n$.

System of those quadratic equations can be rewritten as Because this is a system of quadratic equations, we can have up to 2^p solutions.

Moreover total number of $(p+1)$-element subsets out of n is $\binom{n}{p+1}$. This means that the total number of solutions can be up to $\binom{n}{p+1}2^p$. We denote set of all these solutions H_p.

So if set H_p satisfies 2.102 then it can be used for finding $\mathcal{H}_p$ subset, thus we would be able to use this set for the multidimensional version of the algorithm.

Let $\circ \in +, -$ be either operation $+$ or $-$. Then system of quadratic equations (2.103) is equivalent to 2^p linear systems

$$\begin{aligned} (\boldsymbol{x}_{i_1} \circ_1 \boldsymbol{x}_{i_2})^T \boldsymbol{w} &= y_{i_1} \circ_1 y_{i_2} \\ \vdots \qquad & \qquad \vdots \\ (\boldsymbol{x}_{i_p} \circ_p \boldsymbol{x}_{i_{p+1}})^T \boldsymbol{w} &= y_{i_p} \circ_p y_{i_{p+1}}. \end{aligned} \tag{2.104}$$

Moreover (2.104) is equivalent to the set of equations

$$\begin{aligned}(\boldsymbol{x}_{i_1} \circ_1 \boldsymbol{x}_{i_2})^T \boldsymbol{w} &= y_{i_1} \circ_1 y_{i_2} \\ \vdots \qquad &\vdots \\ (\boldsymbol{x}_{i_1} \circ_p \boldsymbol{x}_{i_{p+1}})^T \boldsymbol{w} &= y_{i_1} \circ_p y_{i_{p+1}}.\end{aligned} \tag{2.105}$$

Elements of set $\mathcal{H}_{\surd}$ then satisfies $r_{i_1^2}(\boldsymbol{w}) = r_{(h:n))}(\boldsymbol{w}) = r_{(h+1:n))}(\boldsymbol{w})$.

Set H_p was proven to be suitable for finding $\mathcal{H}_{\surd}$ in [21] thus that set $\mathcal{H}_{\surd}$ satisfies (2.102) if following assumptions are satisfied (we already assume that 42 holds). First is that for all $\boldsymbol{w}$ $r_{(h:n))}(\boldsymbol{w}) > 0$, second , for all $(\circ_1, \circ_2 \ldots, \circ_{n-1}) \in \times_{i=1}^{n} +, -$ the matrix system of $n-1$ equations

$$\begin{aligned}(\boldsymbol{x}_{i_1} \circ_1 \boldsymbol{x}_{i_2})^T \boldsymbol{w} &= y_{i_1} \circ_1 y_{i_2} \\ \vdots \qquad &\vdots \\ (\boldsymbol{x}_{i_1} \circ_{n-1} \boldsymbol{x}_{i_n})^T \boldsymbol{w} &= y_{i_1} \circ_p y_{i_n}..\end{aligned} \tag{2.106}$$

The seconds assumption prevents the set $\mathcal{H}_{\surd}$ to be empty thus the case where all systems 2.105 are dependent.

This assumption can cause problem because when we assume model with the intercept, thus for all rows $\boldsymbol{x_i}$ of matrix $\boldsymbol{X}$ is $x_{i1} = 1$ then in the case $\circ_1, \circ_2 \ldots, \circ_{n-1} = -$, first column of matrix $\boldsymbol{X}$ only contains zeroes. We can modify this assumption so that instead of using $(\circ_1, \circ_2 \ldots, \circ_{n-1}) \in \times_{i=1}^{n} +, -$ we use only $(\circ_1, \circ_2 \ldots, \circ_{n-1}) \in \times_{i=1}^{n} +, - \setminus (-, \ldots, -)$ [21].

Given that we now know that set H_p and $\mathcal{H}_p$ are suitable for our algorithm and we also know how to calculate them, so let us describe the multidimensional version of the algorithm.

write the pseudocode

Let us now describe the time complexity of this algorithm.

write the time complexity

Chapter 3

Experiments

3.1 Data

3.2 Results

3.3 Outlier detection

In this chapter we describe implementation of all algorithms as well as data sets which we use for experiments. Then a bunch of experiments will be presented. Right now we are finishing implementation of algorithms so we do not have any experimental results yet. This will change this or next week... Chapter will be about 4 pages long.

Conclusion

To date, multiple exact and probabilistic algorithms for calculating LTS estimate has been proposed. Those algorithms have been proposed across the last few decades. Most recently proposed algorithms are just a few years old. That means that research in the filed of LTS algorithms is still prevalent.

Although to date exact finite algorithm has polynomial time, we showed that currently used probabilistic algorithm provide sufficiently fast solutions which, even though are not exact, are good enough. Though it is proven that the exact solution cannot be obtained faster than in polynomial time, we showed that currently used algorithms could be combined to obtain better results.

Algorithms for calculating the LTS estimate is an exciting topic for further research. One of the ways is the research of combining exact algorithms with probabilistic ones. As our experimental results prompt, we would possibly propose probabilistic algorithms which provide even better solutions at the same time.

Another way would be to use algorithms for computing LTS estimate on different robust statistic methods because some of them are very similar to LTS problem. Just put there are still a lot of future works on least trimmed squares, and we are planning to research this in the future.

Bibliography

[1] Rousseeuw, P.; C. van Zomeren, B. Unmasking Multivariate Outliers and Leverage Points. *Journal of The American Statistical Association - J AMER STATIST ASSN*, volume 85, 06 1990: pp. 633–639, doi:10.1080/01621459.1990.10474920.

[2] Massart, D. L.; Kaufman, L.; et al. Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta*, volume 187, 1986: pp. 171–179.

[3] Agulló, J. New algorithms for computing the least trimmed squares regression estimator. *Computational Statistics & Data Analysis*, volume 36, no. 4, 2001: pp. 425–439.

[4] Bernholt, T. Robust estimators are hard to compute. Technical report, Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion, 2006.

[5] Klouda, K. *Studium senzitivity odhadu metodou nejmenších usekaných čtverců*. Bachelor's thesis, Technical University of Prague, Faculty of Nuclear Sciences and Physical Engineering, Department of Mathematics, 2006.

[6] Hawkins, D. M. The feasible solution algorithm for least trimmed squares regression. *Computational statistics & data analysis*, volume 17, no. 2, 1994: pp. 185–196.

[7] Bai, E.-W. A random least-trimmed-squares identification algorithm. *Automatica*, volume 39, no. 9, 2003: pp. 1651–1659.

[8] Rousseeuw, P. J.; Leroy, A. M. *Robust regression and outlier detection*. John wiley & sons, 1987.

[9] Hawkins, D. M.; Olive, D. J. Improved feasible solution algorithms for high breakdown estimation. *Computational statistics & data analysis*, volume 30, no. 1, 1999: pp. 1–11.

[10] Strassen, V. Gaussian elimination is not optimal. *Numerische mathematik*, volume 13, no. 4, 1969: pp. 354–356.

[11] Coppersmith, D.; Winograd, S. Matrix multiplication via arithmetic progressions. *Journal of symbolic computation*, volume 9, no. 3, 1990: pp. 251–280.

[12] Ballard, G.; Benson, A. R.; et al. Improving the numerical stability of fast matrix multiplication. *arXiv preprint arXiv:1507.00687*, 2015.

[13] Anderson, E.; Bai, Z.; et al. *LAPACK Users' Guide*. Philadelphia, PA: Society for Industrial and Applied Mathematics, third edition, 1999, ISBN 0-89871-447-8 (paperback).

[14] Krishnamoorthy, A.; Menon, D. Matrix inversion using Cholesky decomposition. In *2013 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, IEEE, 2013, pp. 70–72.

[15] Hammarling, S.; Lucas, C. Updating the QR factorization and the least squares problem. 2008.

[16] Rousseeuw, P. J.; Driessen, K. V. An Algorithm for Positive-Breakdown Regression Based on Concentration Steps. In *Data Analysis: Scientific Modeling and Practical Application*, edited by M. S. W. Gaul, O. Opitz, Springer-Verlag Berlin Heidelberg, 2000, pp. 335–346.

[17] Hoare, C. A. Algorithm 65: find. *Communications of the ACM*, volume 4, no. 7, 1961: pp. 321–322.

[18] Atkinson, A. C.; Weisberg, S. Simulated annealing for the detection of multiple outliers using least squares and least median of squares fitting. *Institute for Mathematics and Its Applications*, volume 33, 1991: p. 7.

[19] Bartlett, M. S. An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*, volume 22, no. 1, 1951: pp. 107–111.

[20] Hofmann, M.; Kontoghiorghes, E. J. Matrix strategies for computing the least trimmed squares estimation of the general linear and sur models. *Computational Statistics & Data Analysis*, volume 54, no. 12, 2010: pp. 3392–3403.

[21] Klouda, K. An exact polynomial time algorithm for computing the least trimmed squares estimate. *Computational Statistics & Data Analysis*, volume 84, 2015: pp. 27–40.

Appendix A

Data sets

GUI Graphical user interface

XML Extensible markup language

Appendix B

Contents of enclosed CD

- `readme.txt` ..... the file with CD contents description
- `exe` ..... the directory with executables
- `src` ..... the directory of source codes
  - `wbdcm` ..... implementation sources
  - `thesis` ..... the directory of LaTeX source codes of the thesis
- `text` ..... the thesis text directory
  - `thesis.pdf` ..... the thesis text in PDF format
  - `thesis.ps` ..... the thesis text in PS format