KU Leuven

Project

Robust maximum association estimators

Robust Statistics

[G0B16a]

Group 4

Gabriel Bénédict
gabriel.benedict@student.kuleuven.be - r0692805

Konstantina Chatzikonstantinidou
konstantina.chatzikonstantinidou@student.kuleuven.be - r0699766

Gilles Goemaere
gilles.goemare@student.kuleuven.be - r0579273

Bert Goemans
bert.goemans@student.kuleuven.be - r0587061

Dries Van der Plas
dries.vanderplas@student.kuleuven.be - r0583887

Supervised by Prof. Dr. Peter Rousseeuw
peter.rousseeuw@kuleuven.be

November 30, 2018

Contents

1 Introduction **1**

2 Description of the Method **1**
2.1 4 Association Measures 1
2.2 Association Measures Properties 3
2.3 Alternate Grid Algorithm 7

3 Application to Linear Regression **8**

4 Movie Data Analysis **12**
4.1 Comparison and results 13

5 Simulation Study **14**
5.1 First Setting 14
5.2 Second setting 16

6 Affine equivariance **19**

7 Conclusion **20**

Appendix A Additional Mathematical Expressions **22**
A.1 Weighting Vectors' IF 22
A.2 Asymptotic Variances 22

Appendix B Script **24**
B.1 Linear Regression 24
B.2 Movie Data 27
B.3 Simulation Study 33

List of Figures

1 Influence functions of the Pearson (a), Spearman (b), Kendal (c) and M (d) association measures, with $\rho = 0.5$ (Alfons et al., 2016, p.9) 5
2 Asymptotic efficiencies of the weighting vector $\boldsymbol{\alpha}$ in function of the maximum correlation for the four different projection indexes (Alfons et al., 2017) . . . 7
3 The residual plot of the dataset delivery coming from `ltsReg`. 9
4 The residual plot of the dataset delivery coming from `ccaGrid` using Kendall(upper left), Spearman (upper right), M (lower left) and Pearson (lower right). . . . 10
5 The residual plots of the dataset education. The plot on the left comes from LTS model and and the plot on the right from the robust maximum association 11
6 The residual plots of the dataset hbk. The plot on the left comes from LTS model and and the plot on the right from the robust maximum association . 11
7 The residual plots of the dataset telef. The plot on the left comes from LTS model and and the plot on the right from the robust maximum association . 11
8 The residual plots of the dataset wood. The plot on the left comes from LTS model and and the plot on the right from the robust maximum association . 12
9 Boxplots of the prediction errors for both data sets. Blue boxplots (a) correspond to the original data set while white boxplots (b) correspond to the reformed data set. 13
10 Estimated regression coefficients in function of the fraction of outliers when using the first setting with the M maximum association measure. 15
11 Estimated weighting vectors in function of the fraction of contamination with p=3: first component in black, second in red and third in green. The estimation is done by Pearson (left upper), Spearman (right upper), Kendall (left lower) and M (right lower). 17

List of Tables

1 Median fraction of outliers before an estimate of the weighting vectors has a negative component. 15
2 Median fraction of outliers before the estimate of the first component of the weighting vector is lower than 0.5. 19

1 Introduction

In this paper, which is largely based on Alfons et al. (2017), a method to measure the maximum association between two multivariate variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ is studied. Given $\boldsymbol{X}$ and $\boldsymbol{Y}$, the idea is to project both variables on two univariate variables using different projection vectors. In other words, linear combinations of the components of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are taken to form two new univariate variables. The goal is to choose the projection vectors to maximize a certain bivariate measure of association between the two new variables. This measure is also known as the projection index. Examples of such bivariate association measures are the Pearson or Kendall correlation. More formally, the maximum association measure is given by

$$\rho_R(\boldsymbol{X}, \boldsymbol{Y}) = \max\{R(\boldsymbol{\alpha}^t\boldsymbol{X}, \boldsymbol{\beta}^t\boldsymbol{Y})\}$$

with R a measure of association between univariate variables and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ two projection vectors. A restriction imposed on the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is that their norm has to be equal to 1 in order to get a unique solution. The unit vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which maximize the association are also called the weighting vectors and will be denoted as $\boldsymbol{\alpha}_R$ and $\boldsymbol{\beta_R}$.

In this paper, several measures of association and their properties are studied. Special attention will be given to the robustness properties such as the influence function and the breakdown value. This will be studied theoretically as well as by calculating the maximum association measure for several datasets and by a simulation study. Furthermore a grid algorithm will be introduced to compute the weighting vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ numerically.

2 Description of the Method

2.1 4 Association Measures

As described in the introduction, the maximum association measure between two multivariate variables is based on a bivariate association measure, R, between two univariate variables $\boldsymbol{\alpha}^t\boldsymbol{X}$ and $\boldsymbol{\beta}^t\boldsymbol{Y}$. For simplification, we will denote these univariate variables as U and V. Four association measures will be compared. The traditional Pearson correlation is opposed to three measures which are supposedly more robust: the Kendall correlation, the Spearman correlation and a variation of the Pearson correlation based on the M estimator for scale instead of on the covariance matrix.

Pearson Correlation

The Pearson correlation between U and V is given by

$$R_P(U,V) = \frac{\operatorname{cov}(U,V)}{\sqrt{\operatorname{var}(U)\operatorname{var}(V)}}.$$

This association measure is commonly used in statistics, but as will be seen soon, the Pearson correlation isn't robust such that the use of other association measures could be recommended in situations where the data is contaminated by outliers.

Spearman Correlation

The Spearman correlation $R_S(U,V)$ is equal to the Pearson correlation applied on the ranks of the data instead of on the raw data:

$$R_S(U,V) = R_p(\operatorname{rank}(U), \operatorname{rank}(V)).$$

Therefore, the Spearman correlation is not that hardly affected by large outliers. For any given outlier, instead of calculating the magnitude of its value, the magnitude of its rank is used. The rank of an outlier is closer to the ranks of the non-outliers, such that the influence of the outlier will be reduced in comparison to the Pearson correlation.

Kendall Correlation

This third bivariate association measure is given by

$$R_K(U,V) = E\left[\operatorname{sign}\left((U_1 - U_2)\left(V_1 - V_2\right)\right)\right]$$

Based on a sample, the estimated value of Kendall's correlation, also known as Kendall τ, can be calculated as

$$\hat{R}_K = \frac{1}{\binom{n}{2}} \sum_{i<j} \operatorname{sign}\left((u_i - u_j)\left(v_i - v_j\right)\right)$$

This formula illustrates how an extreme value can only impact on the sign of $n-1$ terms in the sum of all $\frac{n(n-1)}{2}$ possible combinations. Adding a small fraction of extreme values to existing vectors U and V therefore only marginally changes the observed correlation which indicates that the Kendall correlation is more robust than the Pearson correlation.

M-Association

The fourth association measure relies on the elements of the bivariate Huber's M-scatter matrix $\boldsymbol{C}(U,V)$ in a similar way as the Pearson correlation depends on the covariance matrix, namely:

$$R_C(U,V) = \frac{C_{12}(U,V)}{\sqrt{C_{11}(U,V)C_{22}(U,V)}}$$

in which the M-scatter matrix $\boldsymbol{C}(\boldsymbol{Z})$ of a two-dimensional variable $\boldsymbol{Z} = (U,V)^t$ is defined by

$$\boldsymbol{C}(\boldsymbol{Z}) = E\left[w_2\left((\boldsymbol{Z}-\boldsymbol{\mu})^t\boldsymbol{C}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu})\right)(\boldsymbol{Z}-\boldsymbol{\mu})(\boldsymbol{Z}-\boldsymbol{\mu})^t\right]$$

with

$$\boldsymbol{\mu} = \frac{E\left[w_1\left((\boldsymbol{Z}-\boldsymbol{\mu})^t\boldsymbol{C}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu})\right)\boldsymbol{Z}\right]}{E\left[w_1\left((\boldsymbol{Z}-\boldsymbol{\mu})^t\boldsymbol{C}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu})\right)\right]},$$

$$w_1(d) = \max\left(1, \frac{\chi^2_{2;0.9}}{d}\right), \quad w_2(d) = c\,\max\left(1, \left(\frac{\chi^2_{2;0.9}}{d}\right)^2\right)$$

and c a consistency factor.

2.2 Association Measures Properties

In the previous section, four different association measures were described. They require to satisfy certain conditions: the association should be symmetric and thus independent of the order in which two vectors are compared (i), it should be invariant to affine transformations (e.g. changing the measurement scale) (ii) and it should be bounded between -1 and 1 to ensure the possibility of comparison between different association measurements (iii).

In order to formulate a first theorem, an additional condition is necessary, although this isn't a necessary condition for being an association measure. Say F_ρ is an elliptically, bivariate distribution of $\boldsymbol{\alpha}^t\boldsymbol{X}$ and $\boldsymbol{\beta}^t\boldsymbol{Y}$ with $\rho = R_p\left(\boldsymbol{\alpha}^t\boldsymbol{X}, \boldsymbol{\beta}^t\boldsymbol{Y}\right)$, the Pearson correlation. The fourth condition then demands that the association of F_ρ as a function of ρ is strictly increasing and differentiable (iv). The four conditions are given more formally as follows:

(i) $R(U,V) = R(V,U)$

(ii) $R(aU+b, cV+d) = \text{sign}(ac)R(U,V)$ for all $a,b,c,d \in \mathbb{R}$

(iii) $-1 \leq R \leq 1$

(iv) $\rho \to R\left(F_\rho\right)$ is a strictly increasing and differentiable function.

Fisher Consistency

For $(\boldsymbol{X}, \boldsymbol{Y}) \sim H$ and H an elliptically symmetrical distribution, we denote the weighting vectors $\boldsymbol{\alpha}_R(H)$ and $\boldsymbol{\beta}_R(H)$. For all association measures which satisfy condition (iv), it holds that $\boldsymbol{\alpha}_R(H)$ and $\boldsymbol{\beta}_R(H)$ are identical and consequently by the Fisher consistency of the Pearson correlation, the Fisher consistency can be expressed generally as

$$\boldsymbol{\alpha}_R\,(H) = \boldsymbol{\alpha}_1/\|\boldsymbol{\alpha}_1\| \quad \text{and} \quad \boldsymbol{\beta}_R\,(H) = \boldsymbol{\beta}_1/\|\boldsymbol{\beta}_1\|$$

with $\boldsymbol{\alpha_1}$ and $\boldsymbol{\beta_1}$ the first canonical correlation vectors.

Influence Function

The fact that the weighting vectors are given by the first canonical correlation vectors can also be used to deduce information about the influence function. The influence function (IF) of the maximum association measure represents the infinitesimal effect of a small contamination on the association measure R, given a certain distribution. The following identity holds:

$$IF((\boldsymbol{x}, \boldsymbol{y}), \rho_R, H) = IF((u_1, v_1), R, F_\rho),$$

with $u_1 = \boldsymbol{\alpha}_1^t \boldsymbol{x}$ and $v_2 = \boldsymbol{\beta}_2^t \boldsymbol{y}$, in which $\boldsymbol{\alpha_1}$ and $\boldsymbol{\beta}_1$ are the first canonical correlation vectors. This formula expresses that the IF of the projection index R determines the shape of the IF for the multivariate association measure ρ_R. Consequently, the influence function of ρ_R is bounded if and only if the influence function of R is bounded.

Furthermore, the influence functions of the weighting vectors $\boldsymbol{\alpha}_R$ and $\boldsymbol{\beta}_R$, $\text{IF}((\mathbf{x}, \mathbf{y}), \boldsymbol{\alpha}_R, H)$ and $IF((\mathbf{x}, \mathbf{y}), \boldsymbol{\beta}_R, H)$, can also be written as a function of $IF((u_1, v_1), R, F_\rho)$. However, these expressions contain the derivative of the influence function, such that even bounded influence functions for the maximum association measure can cause unbounded influence functions for the weighting vectors. The exact relation between the influence function of the weighting vectors and of the projection index can be found in the appendix.

Because the influence function of the maximum association vectors and weighting vectors can be related to the influence function of the bivariate association measure, R, it is useful to look at those influence functions. The influence function of the maximum association vectors for all four described association measures can be found in Figure 1, for a bivariate normal distribution with $\rho = 0.50$. The exact formulas can be found in Alfons et al., 2016.

The figure shows that the influence function of the Pearson correlation is unbounded. Consequently, the Pearson correlation is not robust. This will be confirmed by the experiments shown in the following sections. On the other hand, the other three influence function are bounded. Therefore it is guarantied that outliers will only influence the association proportionally to the level of contamination. Moreover, it could be noticed that all four influence

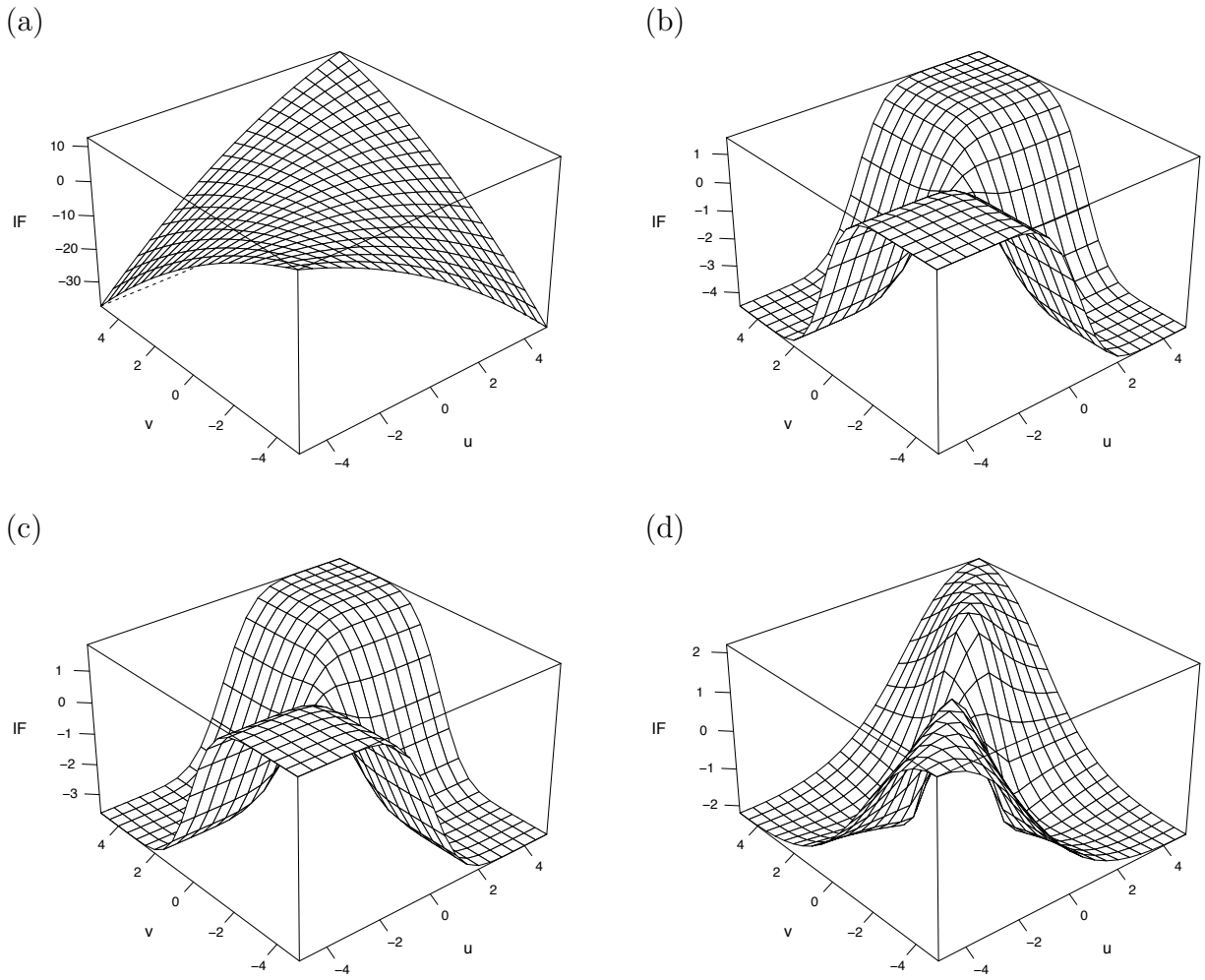

Figure 1: Influence functions of the Pearson (a), Spearman (b), Kendal (c) and M (d) association measures, with $\rho = 0.5$ (Alfons et al., 2016, p.9)

functions are smooth and differentiable which was a necessary condition for the calculation of the influence functions of the weighting vectors.

As a closure remark on the influence functions, we want to point to the fact that the influence function takes only small contamination into account such that other robustness measures such as breakdown point or maximum asymptotic bias could be more informative about non-infinitesimal levels of contamination then the influence function itself.

Asymptotic Variance

In this section, the asymptotic variance of the different association measures for a one dimensional Y-variable will be studied. Given affine equivariance, we can assume that $\Sigma_{\boldsymbol{XX}} = \boldsymbol{I}$, $\Sigma_{\boldsymbol{YY}} = 1$ and $\Sigma_{\boldsymbol{XY}} = (\rho, 0, \ldots, 0)$. Consequently it could be deduced that $\boldsymbol{\alpha}_R = \boldsymbol{\Sigma}_{XY}/\|\boldsymbol{\Sigma}_{XY}\| = (1, 0, \ldots, 0)^t$ and that for elliptically symmetric distributions H_0, the asymptotic variance is a diagonal matrix with diagonal elements equal to

$$ASV\left(\alpha_R, H_0\right)_{jj} = \frac{1}{\rho_R^2 \kappa_R' \left(\rho_R\right)^2} E\left[X_j^2 IF_1^2\left(\left(X_1, Y\right), R, F_\rho\right)\right]$$

if $j > 1$ and 0 otherwise. Given the normality of $H_0 = N(\mathbf{0}, \boldsymbol{\Sigma})$ the asymptotic variances can be calculated for all four bivariate association measures. The exact formulas can be found in the appendix. A plot of the relative asymptotic efficiency can be found in figure 2. This is defined as

$$\mathrm{ARE}\left(\alpha_R, H_0\right) = \frac{ASV\left(\alpha_{R_P}, H_0\right)_{jj}}{ASV\left(\alpha_R, H_0\right)_{jj}}. \tag{1}$$

The asymptotic variance of the Pearson correlation seems to be about 25% higher than the asymptotic variances of the other association measures for all possible values of ρ. Furthermore it could be seen that the relative efficiency of the Spearman correlation is best if $\rho > 0.50$ and becomes better in comparison to the other methods with increasing ρ.

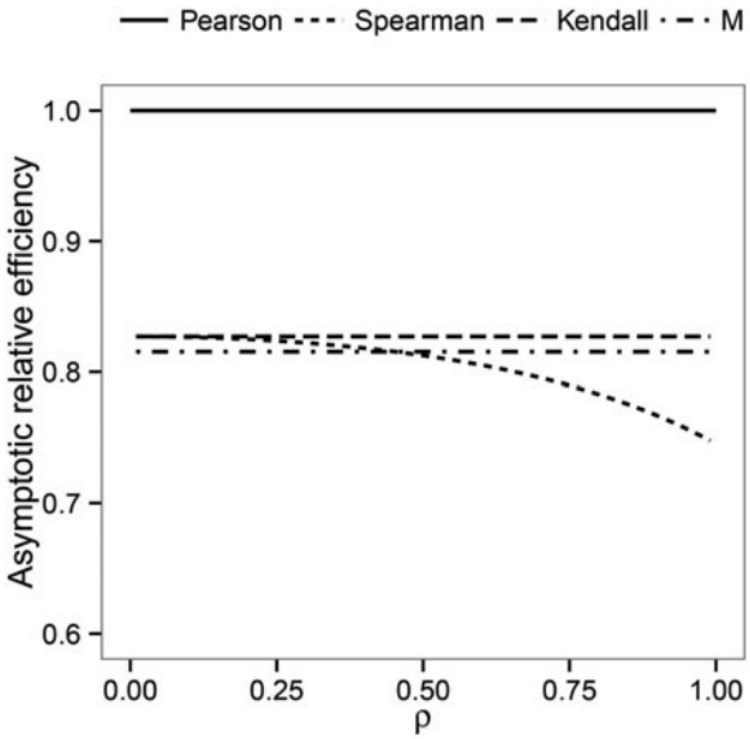

Figure 2: Asymptotic efficiencies of the weighting vector $\boldsymbol{\alpha}$ in function of the maximum correlation for the four different projection indexes (Alfons et al., 2017)

2.3 Alternate Grid Algorithm

Recall that, for multivariate variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ and a bivariate association measure R, the goal is to find projection vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $R(\boldsymbol{\alpha}^t\boldsymbol{X}, \boldsymbol{\beta}^t\boldsymbol{Y})$ is maximal. To numerically calculate the (approximately) correct projection vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, Alfons et al. propose an alternating grid algorithm (2017).

If the dimension of $\boldsymbol{X}$ is equal to 2, then $\boldsymbol{\alpha}$ could be written, because of the normalization, as a vector $(\cos(\theta), \sin(\theta))$, with θ belonging to the interval $[\frac{-\pi}{2}, \frac{\pi}{2})$. If we keep $\boldsymbol{\beta}$ fixed, the problem of finding $\boldsymbol{\alpha}$ can be reduced to a one-dimensional grid search over θ. Take n equidistant grid points in the interval $[\frac{-\pi}{2}, \frac{\pi}{2})$ and evaluate the association measure R in all n points by substituting θ with each grid point and finding the vector for which R is maximal. If n is large enough, the calculated projection vector will be a good approximation of the exact solution.

If the dimension of $\boldsymbol{X}$ is larger than two, a similar approach could be used. For every k^{th} component of $\boldsymbol{\alpha}$, the association measure is maximized in the plane defined by $\boldsymbol{\alpha}$ and $\boldsymbol{e}_k^p$, with $\boldsymbol{e}_k^p$ the k^{th} canonical basis vector. This maximization is done by using the grid search

as described above, but the vector $\boldsymbol{\alpha}$ is now approximated by

$$t_k^P(\boldsymbol{\alpha}, \theta) = \frac{\cos(\theta)\boldsymbol{\alpha} + \sin(\theta)\boldsymbol{e}_k^p}{\sqrt{1 + \sin(2\theta)\boldsymbol{\alpha}_k}}.$$

Call γ the vector of the form described above for which R(γ, β) is maximal. Replace $\boldsymbol{\alpha}$ with γ and repeat this step to update the next component, if this exists.

There exists an algorithm that iteratively performs a series of alternations of the above described steps, where first each component of $\boldsymbol{\alpha}$ is updated while keeping $\boldsymbol{\beta}$ fixed and then vice versa. In each iteration, θ will be restricted to an interval which is half as large as the previous interval. If the improvements are smaller than a given threshold, the algorithm ends and gives back the result. The advantage of this grid algorithm is that it is computationally not very expensive in comparison to a full p-dimensional optimization.

3 Application to Linear Regression

The maximization of the association measure can also be used for performing a linear regression by using a univariate Y. After all, in ordinary least squares regression, the estimated slopes are generated such that the Pearson correlation between the estimated values and the true values is maximal. This idea could now be used to perform linear regression based on the maximal association measures. Indeed, instead of the Pearson correlation one can also try to maximize one of the other association measures. However, for the weighting vectors it was required that the norm of the vector was equal to 1. This is of course not the case for the slopes. Moreover, the method can't be used for the calculation of an intercept. Therefore, the grid algorithm will be applied to the data after a robust standardization of all variables and the response. The obtained $\boldsymbol{\alpha}$ can than be retransformed to get the slopes in the original scale. This method will be illustrated on the following benchmark data sets in the R package robustbase: delivery, education, hbk, telef and wood.

The dataset delivery contains 25 observations giving the time required to service a vending machine in function of the number of products and the distance the deliverer has to walk to the vending machine. The education dataset contains 50 observations and 6 variables giving the per capita expenditure on public education of 50 states in The United States of America and several properties of these states.

The dataset hbk on the other hand is a purely artificial created dataset containing 75 observations in 4 dimensions. It has two groups of outliers. Number 1 to 10 form one group of bad leverage points while 11 to 14 contain good leverage points. The telef dataset contains 24 observations of the total number of international phone calls in Belgium from 1950 to 1973. contrary to the other datasets, this dataset contains only 1 predictor variable. Finally the wood dataset contains 20 observations and 6 variables. It is based on a dataset which was trying to determine the influence of anatomical factors on wood specific gravity from some other variables. However, for the wood dataset some observations were replaced by outliers.

In order to check the performance of the classification of outliers, the residual plots of these datasets coming from the maximum association measures will be compared with those of the robust least trimmed squared (LTS) regression.

To get the results, $\boldsymbol{\alpha}$ was calculated using the alternate grid algorithm from the R package 'ccaPP'. As already discussed, this was done after standardization. The fitted values on the original scale were therefore calculated by standardizing our predictor variables, multiplying with the estimated weighting vector $\boldsymbol{\alpha}$ which results in one value. Afterwards multiplying with the standard deviation of the response variable and finally add the center of the response variable, so

$$\hat{y} = \left(\frac{\boldsymbol{x} - \boldsymbol{\mu}_x}{\sigma_x} \times \boldsymbol{\alpha} \right) \sigma_y + \boldsymbol{\mu}_y.$$

To estimate $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$, σ_x, and σ_y, the robust estimates from `ccaGrid` were used. For the location estimates, this algorithm used the median while for the scale the median absolute deviation was used. Furthermore, the robust distance was calculated to detect leverage points and to be able to make diagnostic plots. The calculation used the Mahalanobis distance based on the robust centre and covariance matrix given by `covMcd`.

The residual plot of the dataset delivery coming from LTS regression, calculated by the `ltsReg`-function, can be found on Figure 3. In Figure 4, the residual plots are shown using the maximum association estimators based on Kendall, Spearman, M and Pearson. Looking at the four plots, there is hardly any difference between the first 3 plots. The plot based on the Pearson correlation is the only one which is remarkably different. As discussed above, the Pearson correlation isn't robust which explains the big differences. The estimated slopes will after all be attracted by the outliers and therefore the Pearson correlation is not a good chose.

When comparing the maximum association diagnostic plots with the LTS diagnostic plot, it seems that almost all of the observations are categorized the same. The only questionable element is 1 which lies on the threshold for being flagged as a vertical outlier on the LTS diagnostic plot in Figure 3 but belongs to the majority of the data in Figure 4.

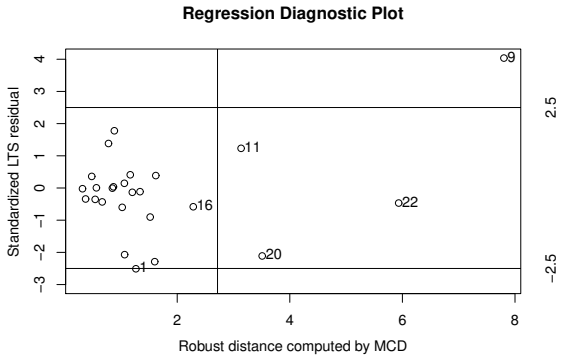

Figure 3: The residual plot of the dataset delivery coming from `ltsReg`.

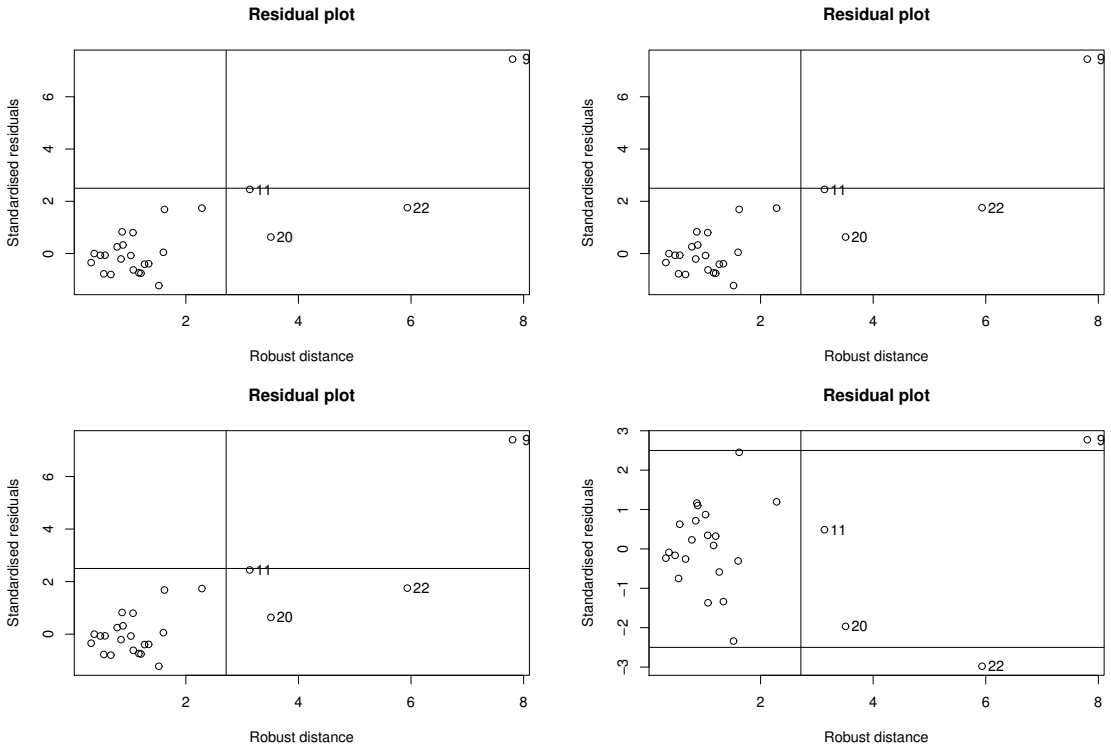

Figure 4: The residual plot of the dataset delivery coming from `ccaGrid` using Kendall(upper left), Spearman (upper right), M (lower left) and Pearson (lower right).

For the other datasets, the diagnostic plots coming from the least trimmed squared regression will only be compared with the plot of `ccaGrid` using the Kendall rank correlation coefficient as association measure because the difference between the Kendall, the Spearman and the M association measures is always very small. The Pearson correlation on the other hand doesn't seems to be useful, as is discussed before. Each time the plot coming from `ltsReg` will be on the left and the regression using the Kendall correlation will be next to it on the right. The results of the other datasets can be seen in figures 5 to 8.

Most of the time each observation is classified the same way both left and right. In Figures 5 and 7, there is in both cases only one element with a different classification. In Figure 5, element 50 which is a bad leverage point is now flagged as a good leverage point. In Figure 7, element 21's classification has changed from a vertical outlier to a regular data point who is almost a vertical outlier. There is also an example in which all elements are classified the same way, namely in Figure 6. Although, also here the magnitude of the residuals of the bad leverage points is clearly lower. This is logical given that all observations have an influence on the Kendall correlation measure.

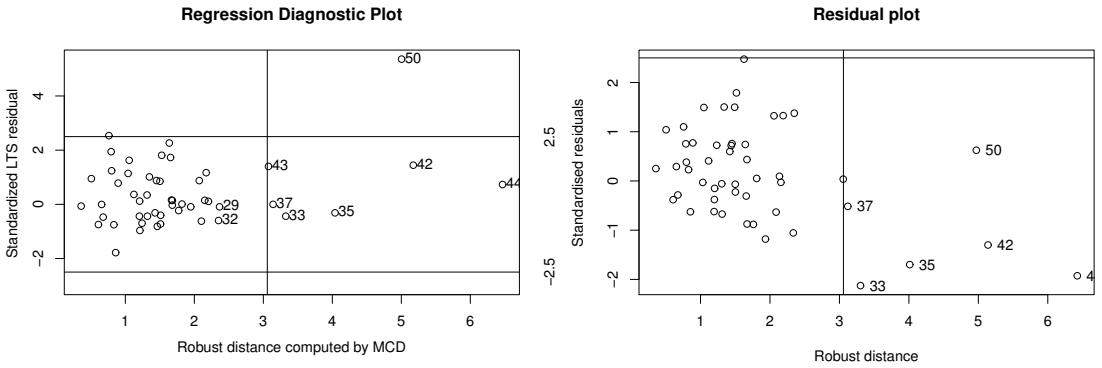

Figure 5: The residual plots of the dataset education. The plot on the left comes from LTS model and and the plot on the right from the robust maximum association

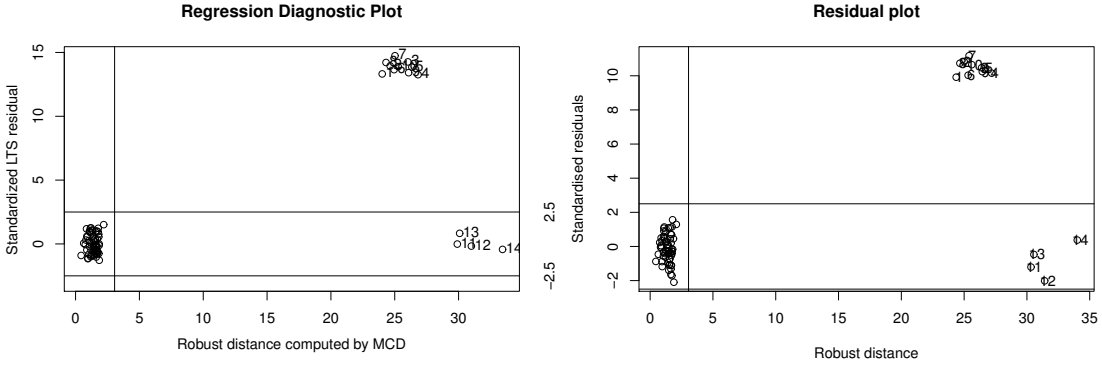

Figure 6: The residual plots of the dataset hbk. The plot on the left comes from LTS model and and the plot on the right from the robust maximum association

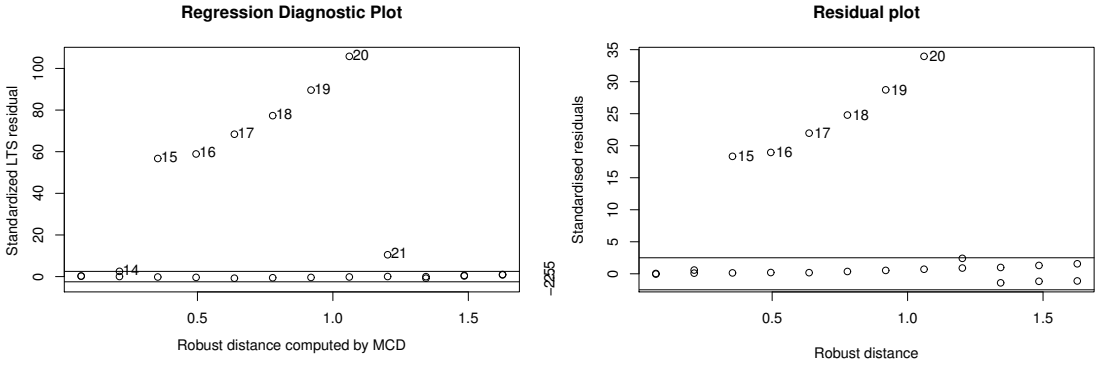

Figure 7: The residual plots of the dataset telef. The plot on the left comes from LTS model and and the plot on the right from the robust maximum association

Finally, the biggest difference in classification is in Figure 8. All the bad leverage points from the LTS model are now classified as good leverage points for the model using the maximum

association measure and element 11 which was a good leverage point on the left picture of Figure 8 is now flagged as a bad leverage point on the right picture. This is probably because there are too many outliers such that our method failed to recognize them. As earlier mentioned the wood dataset contains only 20 elements while the LTS diagnostic plot indicates that 7 of them are outlying, so there is 35% of contamination. The breakdown value of the Kendall correlation is however lower (29.3%) which give raise to the wrong results (Boudt et al., 2010, p.5).

As an overall conclusion, it can be stated that the residual plots coming from `ltsReg` are very similar to the residual plots coming from `ccaGrid` if the ratio of contamination isn't too large. This points to the fact that the robust maximum association measures (Spearman, Kendall and M) can also be used to do linear regression in this case.

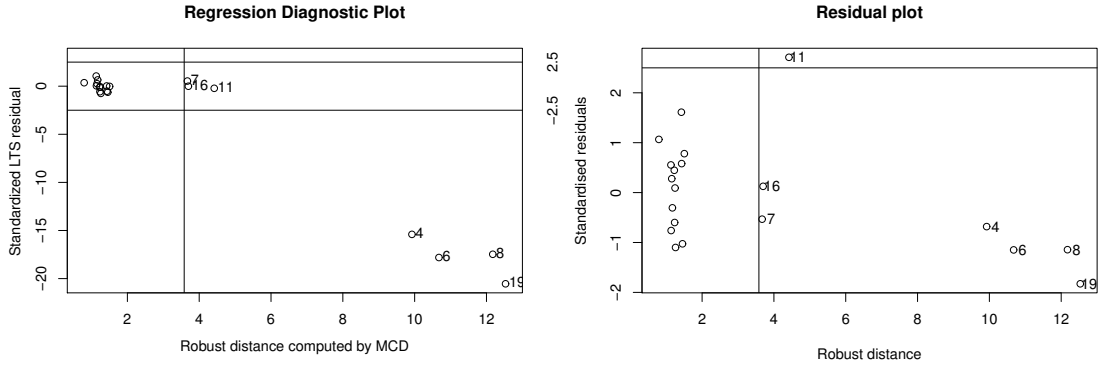

Figure 8: The residual plots of the dataset wood. The plot on the left comes from LTS model and and the plot on the right from the robust maximum association

4 Movie Data Analysis

This part is dedicated to the movies dataset out of the R package 'ggplot2movies'. Again a linear regression will be performed based on maximum association measures. In this dataset, the response variable is the average user rating. First, the analysis was made by using the response and $p = 11$ predictors relevant to the year of release, total budget in U.S. dollars, length in minutes, number of IMDb users who gave a rating, as well as a set of binary variables assessing whether the movie belongs to the genre action, animation, comedy, drama, documentary, romance, or short film. Afterwards, the analysis will be repeated using two extra dummy variables corresponding to the two most common levels of the MPAA rating.

The purpose of the analyses is to compare the four association estimators, namely Pearson, Kendall, M and Spearman together with the least-square regression (LS) and the MM-estimator tuned for 85% efficiency. After removing movies with unknown budget, the dataset contains $n = 5215$ observations. The performance of the linear regression will now be tested by cross-validation. Therefore, the observations are divided into a training and a test set where the test set contains $m = \lfloor n/3 \rfloor$ observations. This process was repeated 1000 times.

The prediction error was calculated for each estimator by Spearman's footrule distance,

$$d = \frac{1}{m}\sum_{i=1}^{m}|r_i - \hat{r}_i|$$

where r_i are the ranks of the movies according to the average user rating and $\hat{r}_i$ are the predicted ranks. Boxplots of the prediction error were constructed for the comparison.

4.1 Comparison and results

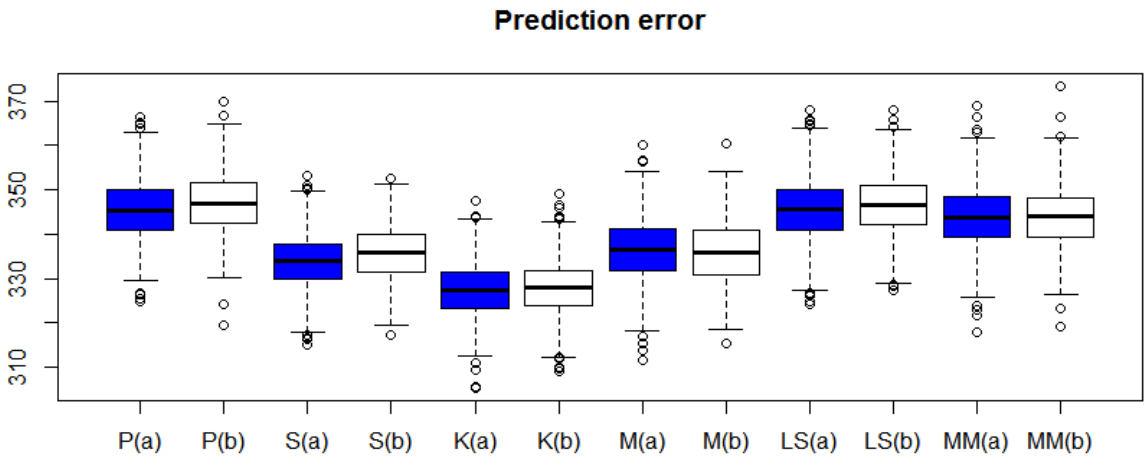

Figure 9: Boxplots of the prediction errors for both data sets. Blue boxplots (a) correspond to the original data set while white boxplots (b) correspond to the reformed data set.

The above boxplots depict the distribution of the prediction error for the six different methods including the 11 predictors (blue boxplots) and using the extra 2 dummy variables (white boxplots). It can be seen that Kendall estimator offers the best result for both data sets, having the lowest prediction error mean. For the original data set, the next most accurate estimator is Spearman's estimator followed by M and MM. Pearson correlation and LS regression estimator which basically lead to the same result, as discussed earlier, are the last and have the highest error mean. So, we can conclude that the robust association estimators perform better than the non-robust ones and especially Kendall's estimator is the most efficient one.

For the reformed data set, after Kendall's estimator, the next most efficient measures appear to be Spearman's and M estimator together, with their means ranging to the same level approximately. MM and Pearson correlation with LS are the ones yielding the highest prediction error.

Overall, there are no big differences between boxplots of the same measure for the two

data sets. In addition, the estimators seem to work the same for the two data sets, meaning that the range of efficiency in terms of lowest prediction error mean is the same both for the original and the data set with the dummies. A possible reason why the addition of two extra predictors does not affect the original results might be that the MPAA ranking, corresponding to the Motion Picture Association of America film rating system to rate a film's suitability for certain audiences based on its content, does not have an important effect on the rating of the movie itself. The foundation of this argument is the fact that a movie which is, for example, suitable for audience over 18 years old does not necessarily influence the opinion of the audience.

5 Simulation Study

In the previous sections, the four different maximum association measures were already used in a regression context and were applied on different datasets. In this section, the methods will be used in a simulation study in order to study the robustness of the methods in a controlled environment.

5.1 First Setting

In the simulation study, data will be generated in two different settings. Each setting starts by generating 200 observations $\boldsymbol{x}_i$ out of a multivariate standard normal distribution with p dimensions. The corresponding responses y_i are given by $y_i = \boldsymbol{\theta}^T \boldsymbol{x}_i + e_i$ with in the first setting $\boldsymbol{\theta} = (1, 1, \ldots, 1)$ and e_i a noise term which is generated out of a normal distribution with mean zero and variance equal to 0.1.

In this setting, a fraction ϵ of the generated points will be turned into bad leverage points by adding 10 to all coordinates of $\boldsymbol{x}_i$ and generating the responses as $-\boldsymbol{\theta}^T \boldsymbol{x}_i^\star$ with $\boldsymbol{x}_i^\star$ the adapted values of $\boldsymbol{x}_i$. Notice that the slopes of the outliers are negative in every dimension while the slopes of the non-outliers are positive. Therefore, the fraction of outliers necessary to make the estimated slopes negative could be used as a measure for the robustness of the regression based on a maximum association measure.

An example of this principle can be seen in Figure 10. This figure shows the estimated regression coefficients by using the M maximum association measure with $p = 5$. The fraction of outliers is step-wise increased in steps of one percent. When the fraction of outlier contamination is small, the estimated regression coefficients are as expected all positive and of similar size. When the fraction exceeds 0.1, a transition phase happens in which some of the estimates are already negative while others are still positive. Increasing the fraction of outliers further causes the estimates to tend to the expected weighting vectors of the outliers (negative and of similar size).

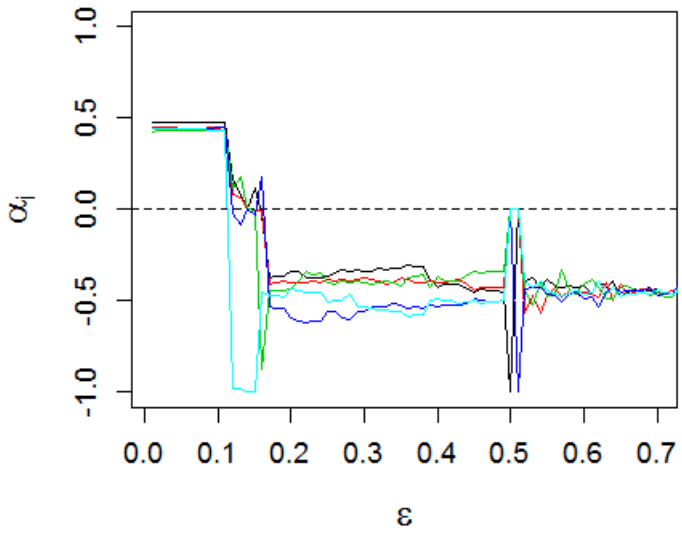

Figure 10: Estimated regression coefficients in function of the fraction of outliers when using the first setting with the M maximum association measure.

As described above, the different maximum association measures can be compared based on the point for which one of the estimated slope becomes negative for the first time. This point could be seen as the breakdown value of the method. To avoid that the results are biased due to the effect of the random errors, the median value over 20 simulations of this breakdown value will be taken. The results can be seen in Table 1 for p equal to 2, 5 and 10. The standard deviations over the 20 studies are in each case smaller than 0.01 which points to the fact that that the results are relatively stable and therefore useful.

Table 1: Median fraction of outliers before an estimate of the weighting vectors has a negative component.

p \ Method	Spearman	Pearson	M	Kendall
2	0.145	0.01	0.16	0.19
5	0.10	0.01	0.12	0.12
10	0.08	0.01	0.10	0.09

The table indicates two phenomena. The first one is the difference in robustness between the different maximum association measures. The Pearson correlation seems to give the least robust result. After all, if only 1 percent of the data is outlying, which is the minimum amount in this study, one of the estimated components of the weighting vector already becomes negative. This is expected because the Pearson correlation depends on the values of the responses, y_i. A single outlier with a response value which is far from the expected value has therefore a huge effect and can cause the estimation to break down. The other methods on the other hand aren't based on the response values themselves (Kendall and Spearman) or give a small weight to the outlying responses (M) and are therefore able to make a more

robust regression estimation.

In low dimensions the Kendall method seems to perform best followed by the M method and the Spearman method. All of them have a breakdown value between 0.145 and 0.19 for p=2, which is a huge advantage over the Pearson method.

The second phenomenon which could be seen in the table is that the higher the dimension of the observations, the lower the breakdown value. This seems the case for all robust methods, but not all methods seem to suffer equally hard. When the number of dimensions is equal to 10, the difference between the breakdown values of the robust regression methods is even not very different anymore.

A possible reason to explain the reduction in breakdown with increasing dimensions could be due to random effects. The more components are present, the more flexibility and thus the higher the probability that one of them changes sign due to the random errors. This is however not the case. Instead of taking the breakdown value as the fraction of the contamination for which one of the components of the weighting vector becomes negative, one can also look at the point where the average of the components becomes negative. By doing so, one can exclude the random effects. It seems however that the results are very similar to the results shown above. The median difference between the to kinds of breakdown value leads after all to differences which are maximally 0.01. Even after excluding the random effects, there is thus still an influence of the number of dimensions.

5.2 Second setting

Because of the equivariance of the maximum association measures under affine transformations, the first setting can be seen after a rotation and scaling as having data with $\boldsymbol{\theta} = (1, 0, \ldots, 0)$ and bad leverage points in the direction of the first basis vector with slope $(-1, 0, \ldots, 0)$. In the second setting we want the outliers to be positioned and having a slope in a direction perpendicular to the direction of the slope. This data will be generated as follows.

The x-data is again generated out of a multivariate normal distribution but now the Y_i-values are generated with $\boldsymbol{\theta} = (1, 0, \ldots, 0)$, so as the first component of $\mathbf{x}_i$, plus an error term. This error term is generated similarly as in the first setting. The outliers are formed by adding 10 to the second component of $\mathbf{x}_i$ and the corresponding response is -10 times this new second component.

Constructing the data like this causes the ordinary data to have a non-zero, positive slope in the first dimension while the outliers have a non-zero, negative slope in the second dimension. All other slopes are zero. As in the first setting, the influence of the outliers on the weighting vectors will be investigated for different fractions of outliers.

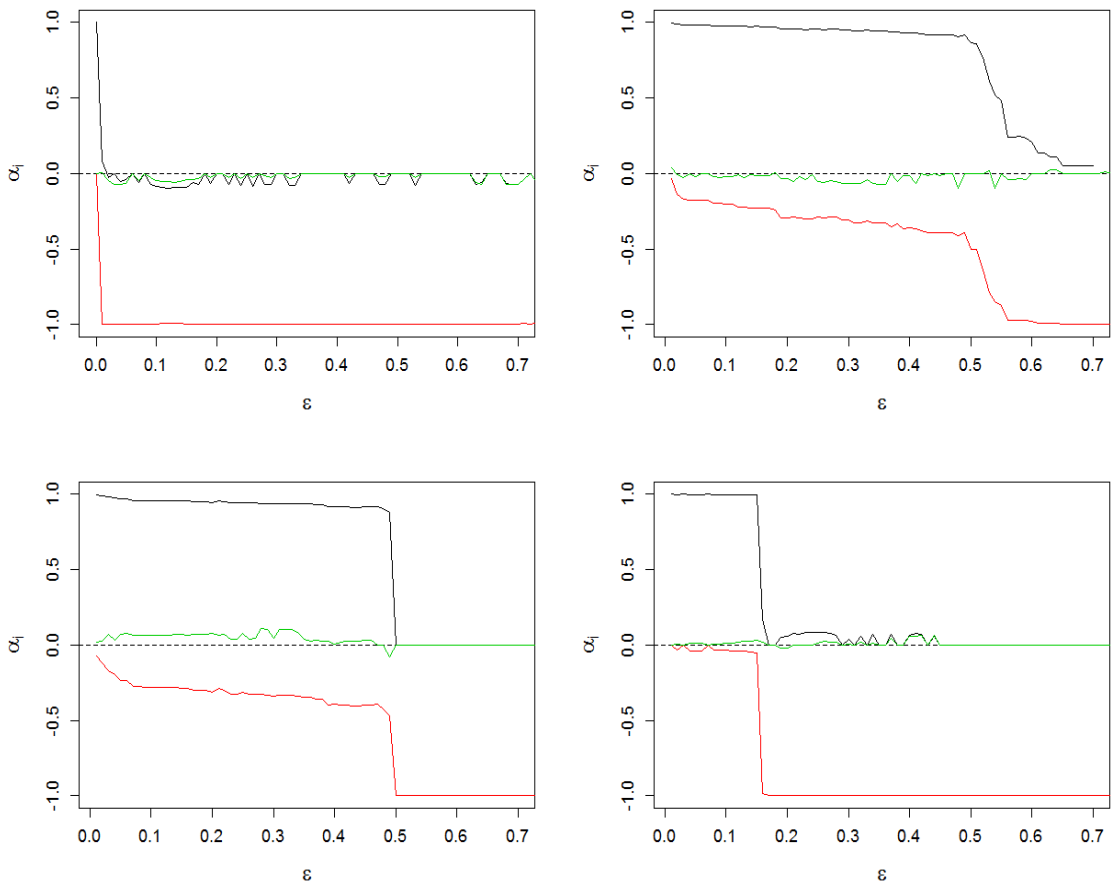

Figure 11: Estimated weighting vectors in function of the fraction of contamination with p=3: first component in black, second in red and third in green. The estimation is done by Pearson (left upper), Spearman (right upper), Kendall (left lower) and M (right lower).

Figure 11 shows plots of the components of the weighting vector in function of the ratio of contamination for the Pearson, Spearman, Kendall and M measures. The figures are generated with a three-dimensional X-space. After all, extra dimensions are expected to act similarly to the third component, which is shown in green. As could be seen, the values of this component are for all methods close to zero, independently of the number of outliers. An explanation is straightforward because both the non-outliers and the outliers have the same zero-slope in this direction.

The result which made use of the Pearson correlation, left upper figure, shows an immediate change of the weighting vector to a vector close to $(0, -1, 0, \ldots, 0)$ which is the expected weighting vectors of the outliers. This is even the case when the number of outliers in minimal. Again this points to the fact that the Pearson correlation isn't robust at all.

When using the Spearman correlation, the estimated first component of the weighting vector changes gradually in the direction of the outlier weighting vector until a contamination

fraction of about 0.50 where the transition occurs faster. Afterwards, the estimated first component evolves again more gradually.

The fact that the transition can occur gradually in the second setting is probably the biggest difference in comparison to the first setting. In the first setting, we expect after all that all components of the weighting vectors are the same in all directions such that the only two possible weighting vectors would have been $\pm\frac{(1,1,\ldots,1)}{|(1,1,\ldots,1)|}$. Therefore we would expect to see both possible weighting vectors in the figure with a clear jump between them when increasing the number of outliers. In the second setting on the other hand, all possible weighting vectors of the form $(a, \pm\sqrt{1-a^2}, 0, \ldots, 0)$ with $|a| < 1$ can occur such that the ratio between the slopes in the first and second direction can change gradually.

For the Kendall association, also a gradually decrease is visible for a low value of outliers. However, at a fraction of outliers of 0.50, the weighting vector becomes suddenly the expected weighting vector of the outliers.

The M-estimator has also a clear transition point but differs in two ways. First, the transition point happens much earlier: already around a contamination fraction of 0.15 the breakdown occurs. Furthermore, there is no gradual decrease before the transition. Instead, the weighting vector stays almost constant in this range of contamination fractions. The M estimator is therefore the most robust in this area but has as disadvantage that it breaks down fast.

We are again interested in the influence of the number of dimensions p. Therefore, a similar table as in the first setting will be generated but the transition point is now determined in another way. In this case it is after all only useful to look at the first (or the second) component of the weighting vector. Moreover, the threshold should be adapted because we expect no transition to negative values for the slope in the first direction. It is chosen to take the threshold of breakdown at $\alpha_1 = 1/\sqrt{2}$, with α_1 the first component of the weighting vector $\boldsymbol{\alpha}$. This value would represent a situation in which the values on the first and second dimension would have a similar magnitude and thus such that the influence of the original data and the outliers is equally important. Here, we expect the magnitude of the components of the weighting vectors in the other directions to be close to zero and thus that they have a negligible effect. The median of the minimum fraction of outliers for which α_1 becomes lower than this threshold over 20 simulations is given in Table 2.

The different values between the association methods were already observed in Figure 11. The Spearman and Kendall method seem to be optimally robust with a breakdown of around 0.50. However, as was seen in the Figures, the breakdown point shouldn't be seen as a clear cut-off point for the Spearman correlation because of the gradually change of the weighting vector.

When comparing the different number of dimensions, there is, in contrast to the first case, no difference visible. The robustness for the second setting is therefore not dimension dependent. This could be explained because the slopes in the third and higher dimensions are always near zero and therefore much smaller than the slopes in the first or second direction.

Consequently they haven't got much influence on the estimation.

Table 2: Median fraction of outliers before the estimate of the first component of the weighting vector is lower than 0.5.

p \ Method	Spearman	Pearson	M	Kendall
2	0.50	0.01	0.16	0.485
5	0.50	0.01	0.16	0.495
10	0.50	0.01	0.16	0.495

6 Affine equivariance

Remember that in the beginning of the previous section, the rotated situation for the first setting with only a slope in one dimension was mentioned. It could be investigated whether the breakdown value would be dimension independent for the same reason as mentioned above. The slopes in the second and higher dimensions would after all be close to zero, both for the outliers as for the non-outliers. Therefore these dimensions would be expected to have no influence on the breakdown value. This would however contradict the affine equivariance of the used method.

A simulation study with the rotated data was done similarly as in the previous settings. This study reveals that the results are indeed dimension independent. The breakdown values are for p equal to 2 similar to the case with p equal to 2 before the rotation: 0.15 for Spearman, 0.01 for Pearson, 0.16 for M and 0.19 for Kendall. In 5 and 10 dimensions, the breakdown values are however, up to small deviations (≤ 0.01), very similar to the case with only 2 dimensions.

This result confirms the reasoning behind the independence of the higher dimensions in the second setting but it seems to contradict the affine equivariance of the maximum association measures. This measures are after all expected to give the same result before and after a rotation.

It could be tested mathematically that the used maximum association measures are themselves equivariant under the rotations. This phenomenon could probably be explained by the fact that the grid algorithm isn't affine equivariant. The algorithm uses after all the canonical basis vectors in the optimization step. Therefore it could be expected that the slopes in the direction of this basis vectors will be treated different from the slopes in other directions such that a different result could be expected after rotation.

7 Conclusion

In this paper, different maximum association measures were studied. Under a mild assumption on the bivariate association measure, it was derived that the weighting vectors are equal to the normalized first canonical correlation vectors. Moreover, it was deduced that the influence function of the maximum association measures and weighting vectors could be related to the influence function of the projection index. Bivariate associations measures with a smooth and bounded influence function such as the Spearman correlation, the Kendall correlation and the M-association measure can therefore be used for robust maximum association estimation. The Pearson correlation on the other hand hasn't got a bounded influence function which leads to non-robust estimations of the maximum association. Moreover it seems that in the case of univariate Y variable, the asymptotic efficiency of the robust maximum association measures is better than those for the Pearson correlation.

In the case of univariate Y, the link between the estimation of the slopes in linear regression and the estimation of the weighting vectors was made. We compared the classification of outliers between the LTS regression model and the linear regression based on the multivariate association measure on several datasets. This revealed that in general the Spearman, Kendall and M associations have a similar classification for outliers if the contamination is not to high. Using the Pearson correlation as association measure on the other hand classified element less similarly than the LTS regression model and so isn't robust at all.

Next, a linear regression on the dataset movies was preformed with 11 variables the first time and 13 variables the second time. The prediction error was calculated using Spearman's footrule. Using Kendall as association measure gives the best result. The difference between the 2 simulations is also very small. A possible explication is that the relation between the 2 extra variables and the response variable is very small or that other variables that are closely related to the 2 extra variables are already in the regression model.

Finally, we tested the robustness of the maximum association measure using the different bivariate association measures in a simulation study. Overall, we can conclude that in most cases using Kendall gives us the best result.

References

Alfons, A., Croux, C., & Filzmoser, P. (2016). Robust maximum association estimators: Technical supplement. supplementary report.

Alfons, A., Croux, C., & Filzmoser, P. (2017). Robust maximum association estimators. *Journal of the American Statistical Association*, *112*(517), 436–445. doi:10.1080/01621459.2016.1148609. eprint: https://doi.org/10.1080/01621459.2016.1148609

Boudt, K., Cornelissen, J., & Croux, C. (2010). The gaussian rank correlation estimator: Robustness properties. K.U.Leuven - Faculty of Business and Economics (Leuven (Belgium)). Retrieved from http://www.econ.kuleuven.be/eng/fetew/int_reports.aspx?group_id=33

Appendices

A Additional Mathematical Expressions

A.1 Weighting Vectors' IF

The influence function of the weighting vectors could be written in function of the weighting vector of the projection index as follows:

$$\mathrm{IF}\left((\mathbf{x},\mathbf{y}),\alpha_R,H\right)=\sum_{k=2}^{p}\frac{1}{\rho_1^2-\rho_k^2}\left\{\mathrm{IF}_1\left((u_1,v_1\right),R,F_\rho)\,\rho_1u_k+\mathrm{IF}_2\left((u_1,v_1\right),R,F_\rho)\,\rho_kv_k\right\}\\ \times\left(I-\frac{\alpha_1\alpha_1^t}{\|\alpha_1\|\|\alpha_1\|}\right)\frac{\alpha_k}{\|\alpha_1\|\kappa_R'\left(\rho_1\right)} \quad (2)$$

and

$$\mathrm{IF}\left((\mathbf{x},\mathbf{y}),\beta_R,H\right)=\sum_{k=2}^{q}\frac{1}{\rho_1^2-\rho_k^2}\left\{\mathrm{IF}_1\left((u_1,v_1\right),R,F_\rho)\,\rho_ku_k+\mathrm{IF}_2\left((u_1,v_1\right),R,F_\rho)\,\rho_1v_k\right\}\\ \times\left(I-\frac{\beta_1\beta_1^t}{\|\beta_1\|\beta_1\|}\right)\frac{\beta_k}{\|\beta_1\|\kappa_R'\left(\rho_1\right)} \quad (3)$$

with for all j: $u_j=\alpha_j^t\mathbf{x}$ and $v_j=\beta_j^t\mathbf{y}$ being the canonical variates for any $(x,y)\in\mathbb{R}^{p+q}$, ρ_i the i^{th} canonical correlation with corresponding canonical correlation vectors $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_i}$. In the formulas IF_1 and IF_2 are respectively the partial derivatives of the influence function in the first and second component.

A.2 Asymptotic Variances

Under the conditions which were given in the paper, formulas could be given for the asymptotic variances for the different association measures. For the Spearman correlation this is

$$ASV(\boldsymbol{\alpha}_R,H_0)_{jj}=\frac{1-\rho^2/4}{\rho^2}16\pi^2E\left[\phi^2(X_1)var(\Phi(\rho X_1+\sqrt{1-\rho^2}Z)|X_1)\right] \quad (4)$$

with Z the standard normal distribution and ϕ the corresponding density function.
For the Kendall correlation this becomes

$$ASV(\boldsymbol{\alpha}_R,H_0)_{jj}=\frac{1-\rho^2}{\rho^2}\frac{2\pi}{3\sqrt{3}} \quad (5)$$

while it is

$$ASV(\boldsymbol{\alpha}_R,H_0)_{jj}=\frac{1-\rho^2}{\rho^2} \quad (6)$$

for Pearson. Finally for the M projection index we get

$$ASV(\boldsymbol{\alpha}_R, H_0)_{jj} = \frac{1-\rho^2}{2\rho^2}\left[E[\gamma^2(d)d] + 0.5E[\gamma(d)\gamma'(d)d^3] + 0.125E[\gamma'(d)^2 d^4]\right] \qquad (7)$$

with

$$d^2 = (u,v)\Sigma_\rho^{-1}(u,v)^t$$

and

$$\gamma(d) = \frac{IF((u,v), R_C, F_\rho)}{IF((u,v), R_P, F_\rho)}.$$

B Script

B.1 Linear Regression

```
library('ccaPP')
library('robustbase')

#Function giving the residual plot for the maximum association measure
Res_plot_1 = function(dataset, col_pred,col_resp,methode ){
  x = as.matrix(dataset[, col_pred])
  y = dataset[, col_resp]
  n = dim(dataset)[1]
  # x contains the predictors and y the responses
  #n is the size of the sample

  #We calculate our model and put it in the variable cca. We call our
   ↪  weighted vector alpha.
  cca = ccaGrid(x,y, method = methode)
  alpha = cca$A

  #Next we calculate the scaled residuals, scaled by mad, and draw a
   ↪  residual plot
  x_cen = x - t(matrix(rep(cca$centerX,n), nrow = length(col_pred)))
  Y_res =
   ↪  y-(((x_cen%*%diag(1/cca$scaleX))%*%alpha)*cca$scaleY+cca$centerY)
  Y_res_mad = Y_res/mad(Y_res)
  #We also calculate the robust distance computed by MCD
  covMcd_x = covMcd(x)
  RD_x = sqrt(mahalanobis(x,covMcd_x$center,covMcd_x$cov ))

  plot(RD_x,Y_res_mad, main = 'Residual plot',xlab = 'Robust
   ↪  distance',ylab = 'Standardised residuals')
  abline(h = 2.5)
  abline(h = -2.5)
  abline(v = sqrt(qchisq(0.975,length(col_pred))))

  outlier = which(abs(Y_res_mad)>=2.5|RD_x
   ↪  >=sqrt(qchisq(0.975,length(col_pred))))
  text(RD_x[outlier]+0.25,Y_res_mad[outlier],labels = outlier)
}
```

```r
#The same function as Res_plot_1 but specific for the telef dataset
Res_plot_telef = function(dataset, col_pred,col_resp,methode ){
  x = as.matrix(dataset[, col_pred])
  y = dataset[, col_resp]
  n = dim(dataset)[1]
  # x contains the predictors and y the responses
  #n is the size of the sample

  #We calculate our model and put it in the variable cca. We call our
  ↪ weighted vector alpha.
  cca = ccaGrid(x,y, method = methode)
  alpha = cca$A

  #Next we calculate the scaled residuals, scaled by mad, and draw a
  ↪ residual plot
   x_cen = x - t(matrix(rep(cca$centerX,n), nrow = length(col_pred)))
  Y_res = y-(((x_cen/cca$scaleX)%*%alpha)*cca$scaleY+cca$centerY)
  Y_res_mad = Y_res/mad(Y_res)
  #We also calculate the robust distance computed by MCD
  covMcd_x = covMcd(x)
  RD_x = sqrt(mahalanobis(x,covMcd_x$center,covMcd_x$cov ))

  plot(RD_x,Y_res_mad, main = 'Residual plot',xlab = 'Robust
  ↪ distance',ylab = 'Standardised residuals')
  abline(h = 2.5)
  abline(h = -2.5)
  abline(v = sqrt(qchisq(0.975,length(col_pred))))

  outlier = which(abs(Y_res_mad)>=2.5|RD_x
  ↪ >=sqrt(qchisq(0.975,length(col_pred))))
  text(RD_x[outlier]+0.05,Y_res_mad[outlier],labels = outlier)
}



#delivery
attach(delivery)
?delivery
head(delivery)
ltsModel = ltsReg(delTime~.,data = delivery, mcd = TRUE)
plot(ltsModel)
Res_plot_1(delivery,c(1,2),3,'kendall')
Res_plot_1(delivery,c(1,2),3,'spearman')
Res_plot_1(delivery,c(1,2),3,'M')
```

```
Res_plot_1(delivery,c(1,2),3,'pearson')
detach(delivery)

#education
attach(education)
?education
head(education)
ltsModel = ltsReg(Y~X1+X2+X3,data = education, mcd = TRUE, use.correction
↪  = FALSE)
plot(ltsModel)
Res_plot_1(education,c(3,4,5),6,'kendall')
detach(education)

#hbk
attach(hbk)
?hbk
head(hbk)
ltsModel = ltsReg(Y~.,data = hbk, mcd = TRUE, use.correction = FALSE)
plot(ltsModel)
Res_plot_1(hbk,c(1,2,3),4,'kendall')
detach(hbk)

#telef
attach(telef)
?telef
head(telef)
ltsModel = ltsReg(Calls~Year,data = telef, mcd = TRUE, use.correction =
↪  FALSE)
plot(ltsModel)
Res_plot_telef(telef,c(1),2,'kendall'))
detach(telef)

#wood
attach(wood)
?wood
head(wood)
ltsModel = ltsReg(y~.,data = wood, mcd = TRUE)
plot(ltsModel)
Res_plot_1(wood,c(1:5),6,'kendall')
detach(wood)
```

B.2 Movie Data

```
# library(ggplot2movies)
library(robustbase)
library(ccaPP)
library(parallel)

#Loading data.
movies=na.omit(movies)
y=movies$rating
x=as.matrix(movies[,c(2,3,4,6,18,19,20,21,22,23,24)])
myseed=set.seed(123)

#Pearson
distance=matrix(0,1000)
for (i in 1:1000){
if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  xtrain=x[train,]
  ytrain=y[train]
  xtest=x[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]
  pearson=CCAgrid(xtrain,ytrain,method='pearson')$A

  pred=xtest%*%pearson
  rank.pred=rank(pred)
  distance[i,1]=mean(abs(rank.pred-rank(ytest)))
}
a=boxplot(distance)

#Spearman
distance1=matrix(0,1000)
for (i in 1:1000){
if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  xtrain=x[train,]
  ytrain=y[train]
  xtest=x[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]
  spearman=CCAgrid(xtrain,ytrain,method='spearman')$A
  pred1=xtest%*%spearman
  rank.pred1=rank(pred1)
```

```
  distance1[i,1]=mean(abs(rank.pred1-rank(ytest)))
}

b=boxplot(distance1)

#Kendall
distance2=matrix(0,1000)
for (i in 1:1000){
if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  xtrain=x[train,]
  ytrain=y[train]
  xtest=x[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]
  kendall=CCAgrid(xtrain,ytrain,method='kendall')$A

  pred2=xtest%*%kendall
  rank.pred2=rank(pred2)
  distance2[i,1]=mean(abs(rank.pred2-rank(ytest)))
}

c=boxplot(distance2)

#M
distance3=matrix(0,1000)
for (i in 1:1000){
if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  xtrain=x[train,]
  ytrain=y[train]
  xtest=x[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]
  M=CCAgrid(xtrain,ytrain,method='M')$A

  pred3=xtest%*%M
  rank.pred3=rank(pred3)
  distance3[i,1]=mean(abs(rank.pred3-rank(ytest)))
}


d=boxplot(distance3)
```

```r
#LS regression.

distancels=matrix(0,1000)
for (i in 1:1000){
  if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
   train=sample(1:5215,3477)
  xtrain=x[train,]
  ytrain=y[train]
  xtest=x[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]

  ls=lm(scale(ytrain)~scale(xtrain)-1,data = movies)$coefficients
  lspred=scale(xtest)%*%ls
  rank.lspred=rank(lspred)
  distancels[i,1]=mean(abs(rank.lspred-rank(ytest)))
}

e=boxplot(distancels,main="LS (a)")

##MM estimator
distanceMM=matrix(0,1000)
for (i in 1:1000){
  if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  xtrain=x[train,]
  ytrain=y[train]
  xtest=x[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]

  con=lmrob.control( estim = "Final",final.alg = "MM",efficiency = 0.85)
  fit2 = lmrob(scale(ytrain) ~scale(xtrain)-1,data=movies,
    method = "MM",cov = ".vcov.w",singular.ok =
     ↪  T,k.max=100000,maxit.scale=100000,control=con)$coefficients
MMpred=scale(xtest)%*%fit2
rank.MMpred=rank(MMpred)
distanceMM[i,1]=mean(abs(rank.MMpred-rank(ytest)))
}

f=boxplot(distanceMM,main="MM (a)")
```

```
table(movies$mpaa)
#Most common levels of mpaa: R & PG-13

#Creating the dummies for the common levels.
dummy=as.numeric(movies$mpaa=='R')
dummy1=as.numeric(movies$mpaa=='PG-13')

#New X
x1=as.matrix(movies1[,c(2,3,4,6,18,19,20,21,22,23,24,25,26)])
movies1=cbind(movies,dummy,dummy1)

#New calculations Pearson.
newdistance=matrix(0,1000)
for (i in 1:1000){
if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  x1train=x1[train,]
  ytrain=y[train]
  x1test=x1[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]
  newpearson=CCAgrid(x1train,ytrain,method='pearson')$A
  newpred=x1test%*%newpearson
  newrank.pred=rank(newpred)
  newdistance[i,1]=mean(abs(newrank.pred-rank(ytest)))
}

newa=boxplot(newdistance)

#New calculations Spearman.
newdistance1=matrix(0,1000)
for (i in 1:1000){
if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  x1train=x1[train,]
  ytrain=y[train]
  x1test=x1[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]
  newspearman=CCAgrid(x1train,ytrain,method='spearman')

  newpred1=x1test%*%newspearman
  newrank.pred1=rank(newpred1)
```

```
  newdistance1[i,1]=mean(abs(newrank.pred1-rank(ytest)))
}

newb=boxplot(newdistance1)

#New calculations Kendall.
newdistance2=matrix(0,1000)
for (i in 1:1000){
if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  x1train=x1[train,]
  ytrain=y[train]
  x1test=x1[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]
  newkendall=CCAgrid(x1train,ytrain,method='kendall')$A
  newpred2=x1test%*%newkendall
  newrank.pred2=rank(newpred2)
  newdistance2[i,1]=mean(abs(newrank.pred2-rank(ytest)))
}

newc=boxplot(newdistance2)

#New calculations M.
newdistance3=matrix(0,1000)
for (i in 1:1000){
if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  x1train=x1[train,]
  ytrain=y[train]
  x1test=x1[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]
  newM=CCAgrid(x1train,ytrain,method='M')$A

  newpred3=x1test%*%newM
  newrank.pred3=rank(newpred3)
  newdistance3[i,1]=mean(abs(newrank.pred3-rank(ytest)))
}

newd=boxplot(newdistance3)


#New calculations LS regression.
```

```
newdistancels=matrix(0,1000)
for (i in 1:1000){
  if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)

  train=sample(1:5215,3477)
  xtrain=x[train,]
  ytrain=y[train]
  xtest=x[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]

  newls=lm(scale(ytrain)~scale(xtrain)-1,data = movies)$coefficients
  newlspred=scale(xtest)%*%ls
  rank.newlspred=rank(newlspred)
  newdistancels[i,1]=mean(abs(rank.newlspred-rank(ytest)))
}

newls=boxplot(newdistancels,main="LS (b)")

##New calculations MM estimator
newdistanceMM=matrix(0,1000)
for (i in 1:1000){
  if (!is.null(seed <- getOption("myseed")))
    set.seed(seed)
  train=sample(1:5215,3477)
  xtrain=x[train,]
  ytrain=y[train]
  xtest=x[setdiff(1:5215,train),]
  ytest=y[setdiff(1:5215,train)]
  newfit2 = lmrob(scale(ytrain) ~scale(xtrain)-1,data=movies,method =
   ↪  "MM",cov = ".vcov.w",singular.ok =
   ↪  T,k.max=100000,maxit.scale=100000,control=con)$coefficients
  newMMpred=scale(xtest)%*%newfit2
  rank.newMMpred=rank(newMMpred)
  newdistanceMM[i,1]=mean(abs(rank.newMMpred-rank(ytest)))
}

newMM=boxplot(newdistanceMM,main="MM (b)")

#All boxplots together.
allin1=data.frame(distance,newdistance,distance1,
                  newdistance1,k$V2,Kendall2_1_$x,distance3,new_M$x,
                  distancels,newdistancels,distanceMM,newdistanceMM)
```

```
j=c(col="blue",col="white",col="blue",col="white",col="blue",col="white",
    col="blue",col="white",col="blue",col="white",col="blue",col="white")

boxplot(allin1,names = c("P(a)","P(b)","S(a)","S(b)","K(a)",
                         "K(b)","M(a)","M(b)","LS(a)","LS(b)",
                         "MM(a)","MM(b)"),main="Prediction
                          ↪  error",ylim(330,390),col=j)
```

B.3 Simulation Study

```
library(ccaPP)
library(mvtnorm)

get.alpha=function(n,p,st,k,method){
#Function estimates the weighting vector alpha for a simulation
#based on n observations of dimension p with the percentage of
#contamination ranging from 0 to 1 in steps of st.
#Two different thetas and contaminations can be selected
#by taking k=1 or k=2. This is as discribed in the assignment.
#The estimation method can be chosen to be:
#'spearman','kendall','pearson' or 'M'.
  x=rmvnorm(n,rep(0,p),diag(1,p,p))
  if (k==1){
    theta=matrix(rep(1,p),1,p)
  }else{
    theta=matrix(c(1,rep(0,p-1)),1,p)
  }
  error=matrix(rnorm(n,sd=sqrt(0.1)))
  y=x%*%t(theta)+error
  a=matrix(0,p,round(1/st))
  b=matrix(0,1,round(1/st))
  for (e in 1:round(1/st)){
    x2=x
    y2=y
    if (k==1){
      x2[1:(n*e*st),1:p]=x[1:(n*e*st),1:p]+matrix(10,(n*e*st),p)
      y2[1:(n*e*st),1]=-(x2%*%t(theta))[1:(n*e*st),1]
    }else{
      x2[1:(n*e*st),2]=x[1:(n*e*st),2]+rep(10,(n*e*st))
      y2[1:(n*e*st)]=-10*x2[1:(n*e*st),2]
    }
    a[,e]=CCAgrid(x2,y2,method=method)$A
  }
```

```
  a
}

result=function(a){
#Function calculates when one of the estimated slopes becomes negative
#for the first time, when the majority of the estimated slopes
#becomes negative for the first time and when the slope in the first
#dimension becomes negative for the first time.
  first.neg=0.01
  for (j in 1:100){
    if(sum(c(a[,j]>0))==p){
      first.neg=first.neg+0.01
    }
    else break
  }
  med.neg=0.01
  for (j in 1:100){
    if(sum(c(a[,j]>0))>p/2){
      med.neg=med.neg+0.01
    }
    else break
  }
  a1.neg=0.01
  for (j in 1:100){
    if(a[1,j]>(1/sqrt(2))){
      a1.neg=a1.neg+0.01
    }
    else break
  }
  list(first.neg=first.neg,med.neg=med.neg,a1.neg=a1.neg)
}

#main program
n=200  #number of observations
p=5    #number of dimensions: 2,5 or 10
st=0.01#increase of outliers in steps of st
k=1    #setting number: 1 or 2
meth='spearman' #methode: 'spearman','kendall','pearson' or 'M'
#calculation of the median and standard deviations of the residuals
res=matrix(0,1,3)
for (i in 1:1){
  a=get.alpha(n,p,st,k,meth)
  res[i,]=unlist(unname(result(a)))
}
```

```
med.res=apply(res,2,median)
names(med.res)=c("first.neg","med.neg","a1.neg")
med.res
apply(res,2,sd)

#plot of alpha in function of fraction of contaminations:
eps=seq(st,1,st)
plot(eps[1:70],a[1,1:70],col=1,type='l',ylim=c(-1,1),ylab=expression(alpha[i]),
    xlab=expression(epsilon),cex.lab=1.5,cex.axis=1.1)
abline(0,0,lty=2)
for (i in 2:p){
  lines(eps,a[i,],col=i)
}
```