Data envelopment analysis with principal component analysis

Toni Huuhka

School of Science

Bachelor's thesis
Espoo 27.8.2020

Supervisor and advisor

Prof. Fabricio Oliveira

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Toni Huuhka

Title Data envelopment analysis with principal component analysis

Degree programme Engineering Physics and Mathematics

Major Mathematics and Systems Sciences **Code of major** SCI3029

Teacher in charge and advisor Prof. Fabricio Oliveira

Date 27.8.2020 **Number of pages** 30 **Language** English

Abstract

The purpose of this bachelor's thesis was to examine how principal component analysis affects the results of data envelopment analysis. When dealing with large amounts of data, the computing power is often a limiting factor in decision making. Therefore, it is important to be able to reduce the the dimensionality of the data without a significant loss of information. Furthermore, the effects of centering and scaling of data prior to the dimensionality reduction were also studied.

First, data envelopment analysis was performed on the entire data set of Scandinavian retailers to compute efficiency scores. Then principal component analysis and robust principal component analysis was used to reduce the number of variables, after which the resulting principal components were used to compute efficiency scores respectively. The effects of centering and scaling of data were studied by performing both versions of principal component analysis using every combination of centering and scaling. The resulting principal components were used in the subsequent data envelopment analysis to compute efficiency scores.

By comparing all sets of results to one another, it became clear that all data must be scaled before the dimensionality reduction in order to produce reliable results. No significant difference between the efficiency scores was found when the data was centered or not centered. It was also observed that the results of data envelopment analysis improved when a more robust version of principal component analysis was used instead of the conventional version.

Keywords Data envelopment analysis, principal component analysis, retail sector, efficiency

Aalto University
School of Science

Aalto-universitetet, PB 11000, 00076 AALTO
www.aalto.fi
Sammandrag av kandidatarbetet

Författare Toni Huuhka

Titel Data envelopment-analys med principalkomponentanalys

Utbildningsprogram Teknisk fysik och matematik

Huvudämne Matematik och systemvetenskaper **Huvudämnets kod** SCI3029

Ansvarslärare Prof. Fabricio Oliveira

Datum 27.8.2020 **Sidantal** 30 **Språk** Engelska

Sammandrag

Syftet med denna kandidatuppsats var att undersöka hur principalkomponentanalys påverkar resultaten av data envelopment-analys. När man hanterar stora mängder data är datorkraften ofta en begränsande faktor i beslutsfattandet. Därför är det viktigt att kunna minska dimensionerna av data utan en betydande förlust av information. Dessutom studerades effekterna av centrering och skalning av data före dimensionsreduceringen.

Först utfördes data envelopment-analys på hela datamängden av skandinaviska återförsäljare för att beräkna effektivitetspoäng. Därefter användes principalkomponentanalys och robust principalkomponentanalys för att minska antalet variabler, varefter de resulterande principalkomponenterna användes för att beräkna respektive effektivitetspoäng. Effekterna av centrering och skalning av data studerades genom att utföra båda versionerna av principalkomponentanalys med användning av varje kombination av centrering och skalning. De resulterande principalkomponenterna användes i sin tur för att beräkna effektivitetspoäng.

Genom att jämföra alla uppsättningar av resultat med varandra blev det klart att all data måste skalas innan dimensionsreduceringen för att producera tillförlitliga resultat. Ingen signifikant skillnad mellan effektivitetspoängen hittades när data var centrerade eller inte centrerade. Det observerades också att resultaten av data envelopment-analys förbättrades när man använde en mer robust version av principalkomponentanalys.

Nyckelord Data envelopment analysis, principalkomponentanalys, detaljhandel, effektivitet

Contents

1 Introduction

Operational efficiency is an important performance indicator in every industry, as it directly affects the magnitude of profit or loss a company makes. Therefore, improving the operational efficiency of a company is a goal for every decision-maker. However, if everyone knew how to improve operational efficiency, many bankrupt companies would still be in business.

As a by-product of technological advancements, there is more data available than ever before. Naturally, this raises the opportunity for *data-driven decision making*, where by looking at data and finding patterns, beneficial conclusions can be made. Modern companies hardly make any major decisions without compelling evidence that the decision is beneficial, although not all decisions are made with the necessary evidence to justify them.

In order for a company to know how well it operates its business compared to other similar businesses, a ranking system with the operational efficiencies of all reference companies could be introduced. The nonlinear programming method *Data Envelopment Analysis* (DEA) introduced by Charnes et al. (1978) was first used to measure the relative efficiency of theoretical *Decision Making Units* (DMUs). The model was later applied to the retail sector by Donthu and Yoo (1998), who studied over several consecutive years the productivity of 24 stores belonging to the same franchise. Mostafa (2009) studied the relative efficiencies of 47 separate retail companies in USA from a financial perspective using DEA to determine the appropriate efficiency scores.

As the amount of data that is available for computation is increasing, so is the need for more computing power. However, currently there is more data available than what can be processed within a desired time frame, which raises the demand for more effective data usage. For this purpose, Adler and Golany (2001) proposed the usage of *Principal Component Analysis* (PCA) in combination with DEA in their study on airline networks to overcome the drawbacks presented by having several variables in the DEA model. This PCA-DEA method has since been adapted to various studies ranging from discrimination in DEA Adler and Yazhemsky (2010) to quality of life in Estonia Põldaru and Roots (2014).

The study conducted for this thesis focuses on measuring the relative efficiency between Scandinavian retailers and observing how the relative efficiency is affected when PCA and *Robust Principal Component Analysis* (RPCA) is performed prior to DEA. The study was initially performed in three stages, where conventional DEA with variable returns to scale was first performed to set the baseline. In the second part, PCA was performed separately on both input and output variables, after which the resulting principal components were used as input and output variables in the DEA calculation. Lastly, the study was conducted by performing RPCA on the original input and output variables and the resulting principal components were used to form the DEA model.

Furthermore, the effects of *variance scaling* and *mean centering* in PCA and RPCA are investigated, as well as how these affect the final results in DEA. This is studied by repeating the PCA-DEA and RPCA-DEA calculations with the additional change of either not scaling the data, not centering the data, or not centering or scaling it prior to the variable reduction technique.

2 Data and methods

2.1 Data

The data used for this thesis was obtained from Orbis; a database by Bureau van Dijk with information on companies from all countries. The chosen data consisted of Scandinavian retailers falling under the industry classification code *47* (*Retail trade, except of motor vehicles and motorcycles*) used in the European Union, with an annual revenue of at least € 50 million.

The data in Orbis is not complete for all companies, which forced us to manually check the exported data and remove all rows with incomplete information. Furthermore, duplicate and semi-duplicate rows are common in Orbis as the database contains information on both the operative companies and their holding companies. These were also removed from the data based on visual examination; for instance, if the holding company was reported with a few employees but with the combined revenue of three subsidiaries and the three subsidiaries were reported separately, the holding company was removed. After all conditions were met, the final data set consisted of 221 observations with five different variables.

According to Donthu and Yoo (1998), measuring retail efficiency should incorporate not only the traditional financial aspect, but also the behavioural aspect of retailing. The behavioural aspect requires information on customer satisfaction and similar variables. However, this sort of information is not available in Orbis, limiting the focus of this study solely on retailing efficiency from the financial aspect. In addition, Donthu and Yoo (1998) argue that profit, perhaps the most intuitive output measure, should not be used as an output variable as it is an aggregate of revenue and cost, defined as price times output quantity and factor price times input quantity, respectively. Therefore, the output variables chosen for this study were *Sales* (Y1) and *Cost of Employees* (Y2). The corresponding input variables were *Number of Employees* (X1), *Current Number of Directors & Managers* (X2), and *Current Number of Advisors* (X3). Summary statistics of the aforementioned variables are presented in Table 1 and the Pearson correlation coefficients between the chosen variables are presented in Table 2.

	Output Variables		Input Variables		
	Y1	Y2	X1	X2	X3
Mean	331.10	43.97	1007	5.81	1.24
Median	122.00	20.00	470	4.00	1.00
Std Dev	757.89	91.62	2224.27	4.47	0.75
Minimum	50.00	1.00	1.00	0.00	0.00
Maximum	7912.00	985.00	27497	31.00	10.00

Table 1: Summary statistics of the variables used in the DEA model.

	Y1	Y2	X1	X2	X3
Y1	1.0000				
Y2	0.9573	1.0000			
X1	0.9173	0.9676	1.0000		
X2	0.4708	0.5245	0.4931	1.0000	
X3	0.1576	0.1733	0.1845	0.2295	1.0000

Table 2: Pearson correlation coefficients between all input and output variables.

A close inspection of the correlation coefficients present very high correlations between outputs Y1 and Y2 ($r = 0.9573$) as well as between input X1 and both outputs Y1 ($r = 0.9173$) and Y2 ($r = 0.9676$).

2.2 Data Envelopment Analysis

DEA is a nonparametric linear programming methodology designed to measure and compare the relative efficiency of a set of decision-making units (DMUs). The model was first proposed by Charnes et al. (1978) as a tool for evaluating decision units, but it has since been applied to various fields, as summarized by Seiford (1997). One of the main assumptions in the DEA model is whether constant returns to scale (CRS) or variable returns to scale (VRS) are used. The original model introduced by Charnes et al. (1978) uses CRS, whereupon the original model is colloquially known as the *CCR* model, whereas the *BCC* model introduced by Banker et al. (1984) uses VRS. We assumed that the returns to scale are more likely to vary than to remain constant in this study, therefore the BCC model was used.

The DEA model is capable of handling multiple inputs and outputs for each DMU, to which an efficiency score is computed as the maximum ratio of weighted outputs to weighted inputs. This ratio is subject to the constraint that similar ratios are less than or equal to one using the aforementioned weights. Thus, the maximum efficiency h_0 for DMU 0 is determined by equations (1) and (2):

$$\max h_0 = \frac{\sum_{i=1}^{n} u_i y_{i0}}{\sum_{j=1}^{m} v_j x_{j0}} \tag{1}$$

$$\text{Subject to } \frac{\sum_{i=1}^{n} u_i y_{ir}}{\sum_{j=1}^{m} v_j x_{jr}} \leq 1 \qquad \forall r = 1, ..., s \tag{2}$$

$$u_i, v_j \geq 0; \;\; i = 1, ..., n; \; j = 1, ..., m$$

where y_{ir} and x_{jr} are the i^{th} output and j^{th} input observations for the r^{th} DMU, and u_i and v_j are the weights to be estimated for the 0^{th} DMU. By applying the functions above to a set of DMUs, DEA identifies and connects the points with the lowest total input for any given total output, creating an *efficient frontier*. The points on the frontier are considered efficient and receive the efficiency score one, whereas all points not on the frontier are considered inefficient and consequently obtain an efficiency score between zero and one. A simplified visualization of the efficient frontier by Hui and Wan (2013) is depicted in Figure 1.

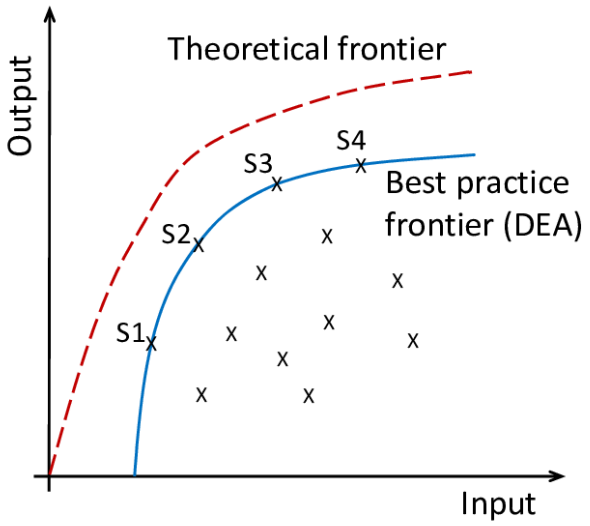


Figure 1: An illustrative drawing of an efficient frontier as computed with DEA.

One of the main drawbacks of DEA is its sensitivity to outliers. Especially when using real-life data, outliers are common and decrease the precision of the DEA as shown by Donthu et al. (2005). However, the authors also note that in a scenario where many efficient units are identified, having a natural outlier or adding a dummy outlier may decrease the number of efficient units, ultimately leading to more precise results from DEA. Another common drawback of the DEA is the tendency of an increasing number of efficient units as the number of variables used in the model increases, as noted by Mostafa (2009).

2.3 Principal Component Analysis

Põldaru and Roots (2014) argue that it is necessary to reduce the impact of outliers by reducing the number of variables in the DEA model. Furthermore, the authors argue that reducing the dimensionality of the DEA structure improves the overall confidence of the DEA. A solution to overcome these issues is introduced by Adler and Golany (2001) by combining PCA and DEA. The underlying theory is achieved by using the original variables to create principal components that are less sensitive to statistical noise than the original variables. Adler and Golany (2001) state that PCA is used to explain the structure of variance in a data set by creating linear combinations of the original variables, and that the new principal components generally account for 80-90% of all variance in the data, in which case the principal components can replace the original variables without a significant loss of information. Figure 2 provides a visualization of the fundamental purpose of performing PCA on a collection of data points.

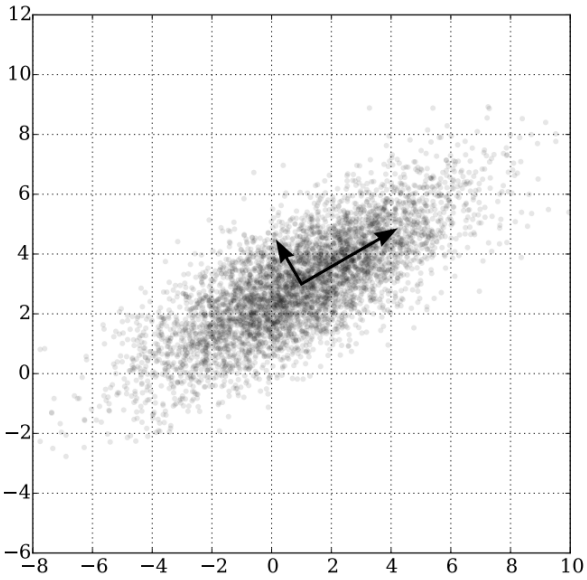


Figure 2: An illustration of the two leading principal components of a multivariate Gaussian distribution (Wikipedia, 2016).

There are alternatives regarding how PCA is performed, the two most common approaches being *Singular value decomposition* (SVD) and *Eigendecomposition*. In this study we decided to use the R function *prcomp*, which uses SVD for computing principal components, for its ease of use and wide selection of plotting possibilities. SVD is a linear algebra technique for decomposing matrices by using a combination of both eigendecomposition and polar decomposition.

As we were using real-life data in this study, outlier values were expected to be present during the analysis. To overcome this issue, PCA was performed on both input and output variables separately, as suggested by Põldaru and Roots (2014). A standard process of PCA is to scale and center the data so that each variable

has unit-variance and the empirical mean of each observation is zero. This is done in order to prevent variables with high absolute values of dominating the variance minimization performed in PCA, as could be expected to happen based on the variables presented in Table 1. The general view is that inputs and outputs have to be strictly positive in DEA, as noted by Adler and Golany (2001). However, the results from PCA are not necessarily positive. To ensure strictly positive data in the DEA model, every principal component was increased by the smallest value in the vector plus one, as defined by equation (3).

$$\tilde{PC}_i = PC + a, \qquad a = \min\{PC_i\} + 1 \tag{3}$$

2.4 Robust Principal Component Analysis

Robust PCA is a widely used modification of conventional PCA that is designed to be less sensitive to outliers and other corrupted observations. There are various alternatives as how to implement a more robust version of PCA, however, we chose to use the *pca* function from the *pcamethods* package, which uses a robust modification of SVD. This version of SVD is achieved by replacing the Euclidean distance with the Manhattan distance, meaning that instead of minimizing the square root of the sum of squares:

$$||x||_2 = \sqrt{x_1^2 + ... + x_n^2}$$

we minimize the sum of absolute values:

$$||x||_1 = \sum_{i=1}^{n} |x_i|$$

As it is designed to perform better with corrupt observations, RPCA was chosen for this study in order to observe whether the usage of RPCA significantly improves the DEA results compared to PCA-DEA. Just as with conventional PCA, principal components are computed separately for inputs and outputs using RPCA, after which the principal components are used in the DEA model.

3 Results

3.1 DEA results

First, we performed DEA on the entire collection of data with 221 DMUs, using all three inputs and all two outputs. The summary statistics representing all three quartiles as well as the minimum and maximum values of the resulting efficiency scores are presented in Table 3.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.3298	0.5140	0.6620	0.6621	0.7696	1.0000

Table 3: Summary statistics of the computed efficiency scores.

A histogram of the efficiency scores is presented in Figure 3. The number of DMUs that were deemed efficient was *13*.

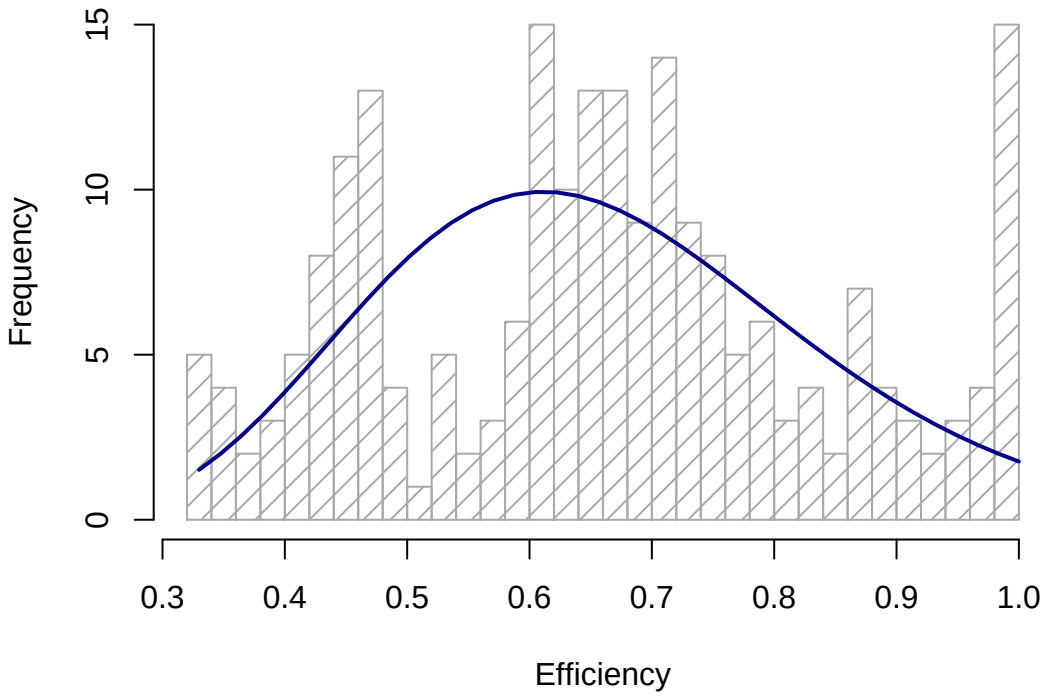


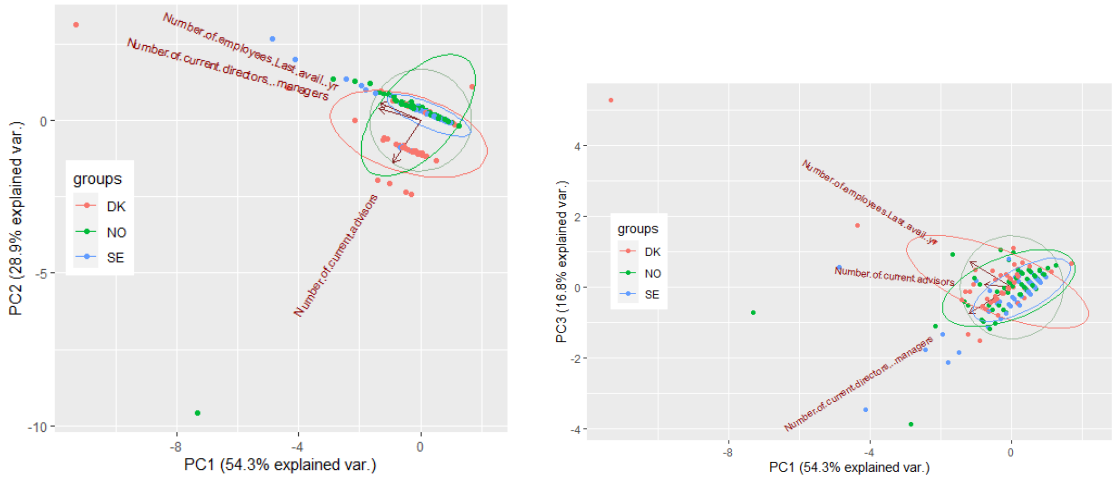
Figure 3: A fitted gamma curve ($\alpha = 13.03$, $\beta = 19.68$) over the frequency distribution of computed efficiency scores.

3.2 PCA-DEA results

	Input Variables			**Output Variables**	
	IPC1	IPC2	IPC3	OPC1	OPC2
Standard Deviation	1.2765	0.9307	0.7102	1.3990	0.2066
Proportion of Variance	0.5432	0.2887	0.1681	0.9787	0.0213
Cumulative Proportion	0.5432	0.8319	1.0000	0.9787	1.0000

Table 4: Summary of the principal components computed with conventional PCA

As can be seen from Table 4, the first two principal components explain roughly 83.2% of the variance in the data, meaning that most of the information and variance has been preserved. In the same manner, the first principal component of the output variables explains roughly 97.9% of the variance whereas the second component only explains the remaining 2.1%, meaning that virtually no extra information is provided by the second component. The high degree of explanatory power of the first component does not come as a major surprise, as the two output variables were highly correlated as stated in Table 2.



(a) The first two principal components plotted against one another.

(b) The first and third principal components plotted against one another.

Figure 4: Visualization of the computed principal components.

Figure 4 shows a scatter plot of the first two principal component scores plotted against one another, where each point is grouped according to the country of origin. The data points form clear lines in the scatter plot, which indicates that outliers are present in the data, as they contribute with a significant amount of variance compared to other data points. A common way of identifying outliers is to plot the residuals as multiples of the standard deviation, where a cut-off point has been predetermined. This is colloquially known as the criterion of "being more than X standard deviations away from the mean". However, because the computed principal components indicate that the data contains significant outliers a more robust estimator was used to detect outliers. By creating a graph of residuals as multiples of the *median absolute deviation*, the outliers are clearly shown.

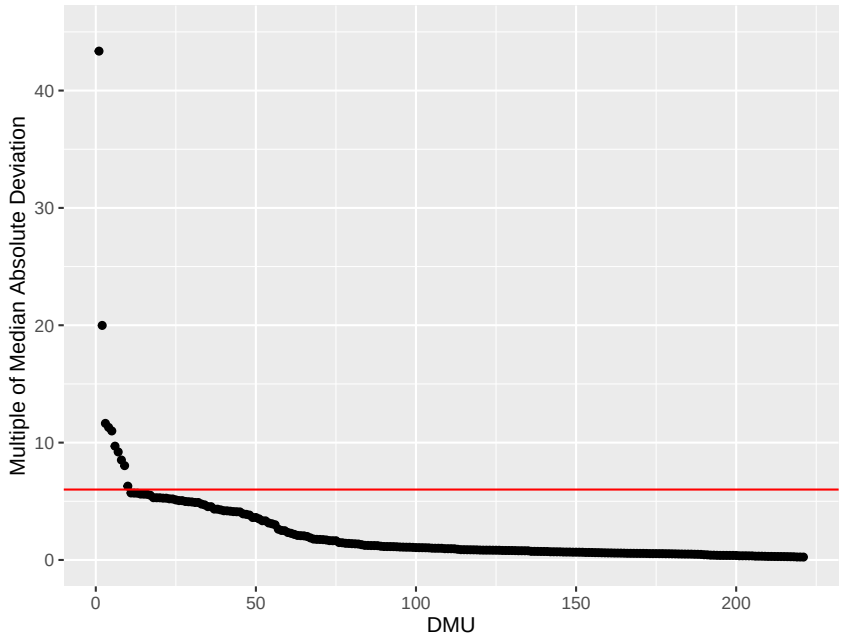


Figure 5: Visual representation of each DMU's distance to the median as a multiple of the median absolute deviation. The red line indicates six median absolute deviations from the median.

Figure 5 shows that there are several DMUs whose computed principal components deviate significantly from the median values. The magnitude by which they deviate is sufficient to draw the leading principal components towards themselves, meaning that the PCA calculation mistakenly considers the variance from the outlier points as signal instead of noise. Thus, the resulting principal components are affected to some extent by the outliers, which is possibly affecting the subsequent DEA calculation.

We used the first two principal components (IPC1 and IPC2) as input variables and the first principal component (OPC1) as the sole output variable in the DEA model. The summary statistics of the resulting efficiency scores are presented in Table 5.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.3594	0.6990	0.8299	0.7742	0.8910	1.0000

Table 5: Summary statistics of the computed efficiency scores.

A histogram of the efficiency scores visualizing the distribution is presented in Figure 6. The number of efficient DMUs in this part was *8*.

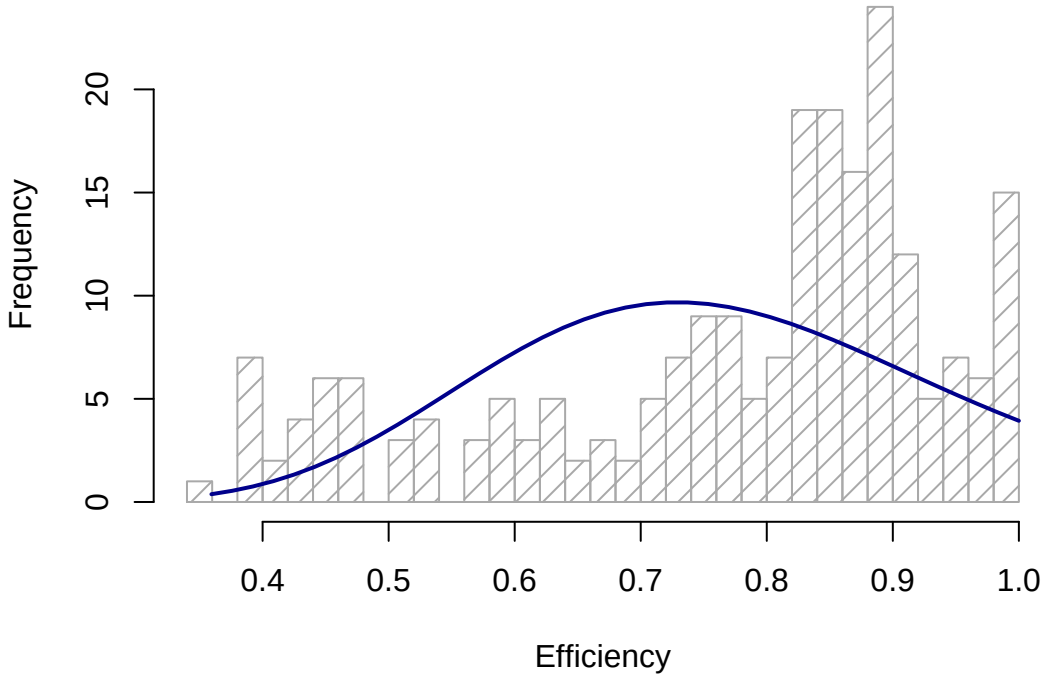


Figure 6: A fitted gamma curve ($\alpha = 17.19$, $\beta = 22.21$) over the frequency distribution of computed efficiency scores.

3.3 RPCA-DEA results

	Input Variables			**Output Variables**	
	IPC1	IPC2	IPC3	OPC1	OPC2
Standard Deviation	1.266	0.977	0.972	1.399	1.000
Proportion of Variance	0.534	−0.087	0.129	0.978	−0.460
Cumulative Proportion	0.534	0.447	0.576	0.978	0.519

Table 6: Summary of the principal components computed with RPCA

As can be seen from Table 6, the first principal component explains roughly 53.4% of the variance in the data, but the second principal component lowers the cumulative variance explained by roughly -8.7%. If a principal component explains a negative portion of the variance, it usually means that outliers have affected the calculation to such an extent that the principal component, which is in itself a linear component, does not fit in the data. This can be visually confirmed by examining Figure 7. The

third principal component fits better in the data, explaining an additional 12.9% of the variance.

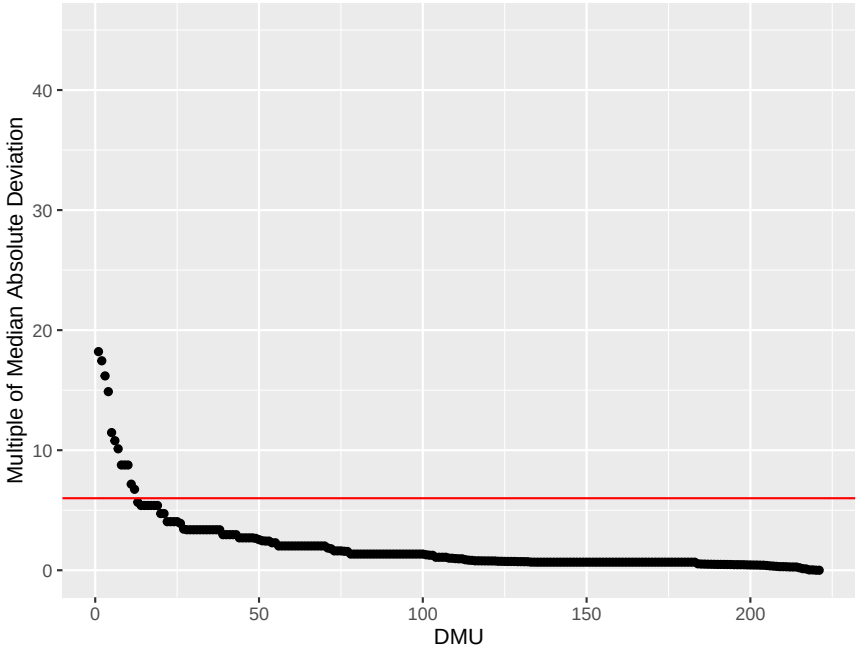
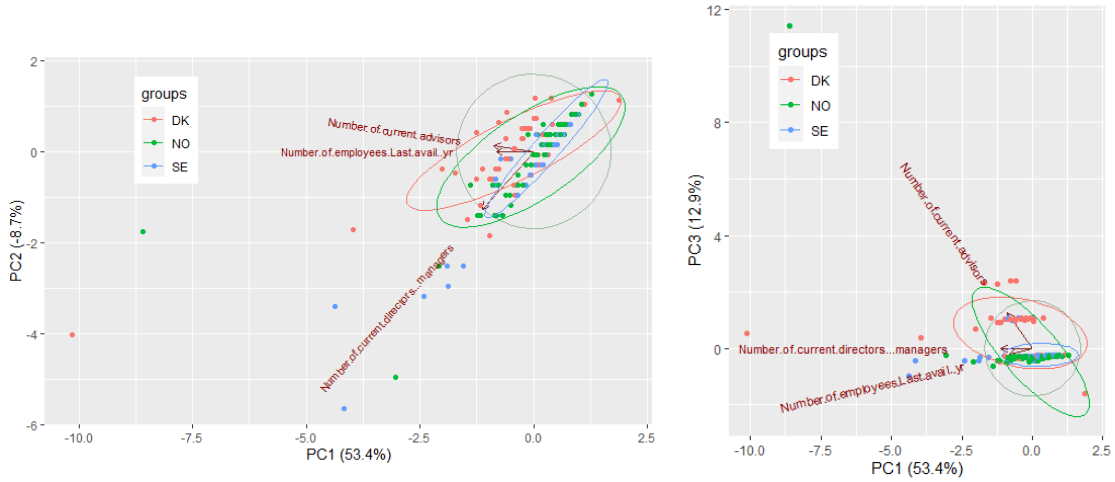


Figure 7: Visual representation of each DMU's distance to the median as a multiple of the median absolute deviation. The red line indicates six median absolute deviations from the median.

As seen with regular PCA, there are several outliers that differ from the median by multiple median absolute deviations. However, the benefit of using RPCA is apparent when comparing Figure 7 to Figure 5, as the outlier observations above the red line do not deviate as much from the rest of the observations as with regular PCA. Nonetheless, the outliers are still significantly affecting the principal components as shown when plotting the principal components against one another. Both Figures 8a and 8b exhibit clear lines as formed by the computed principal component scores.



(a) The first two principal components plotted against one another.

(b) The first and third principal components plotted against one another.

Figure 8: Visualization of the computed robust principal components.

In this part we also used the first two principal components from the original input variables and the first principal component from the output variables in the DEA model. The summary statistics of the resulting efficiency scores are presented in Table 7:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.4063	0.6157	0.7406	0.7186	0.8290	1.0000

Table 7: Summary statistics of the computed efficiency scores.

A histogram of the efficiency scores visualizing the distribution is presented in figure 9. The number of efficient DMUs in this part was *9*.

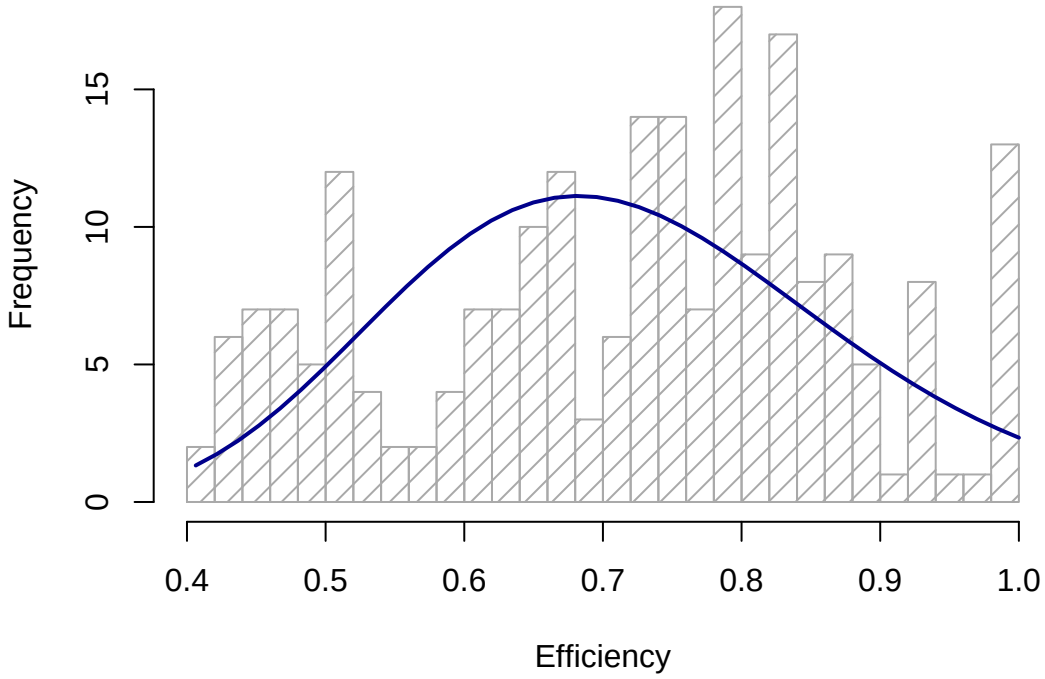


Figure 9: A fitted gamma curve ($\alpha = 19.69$, $\beta = 27.40$) over the frequency distribution of computed efficiency scores.

3.4 Non-centered PCA-DEA

As both PCA and DEA are sensitive to outliers, we would like to investigate whether scaling and centering has an effect on the obtained DEA results. By performing both versions of PCA on the original input and output variables separately first without centering, then without scaling, and lastly without scaling or centering, we will determine the optimal combination to use. The summary statistics of the resulting principal components of non-centered data are presented in Table 8.

	Input Variables			**Output Variables**	
	IPC1	IPC2	IPC3	OPC1	OPC2
Standard Deviation	1.4846	0.7597	0.4678	1.4015	0.1894
Proportion of Variance	0.7346	0.1924	0.0729	0.9821	0.0179
Cumulative Proportion	0.7346	0.9270	1.0000	0.9821	1.0000

Table 8: Summary of the principal components computed with PCA without centering of data.

The share of variance explained by the first principal component has increased from 54.3% to roughly 73.5% which is in line with the fact that centering decreases differences in scale between data points.

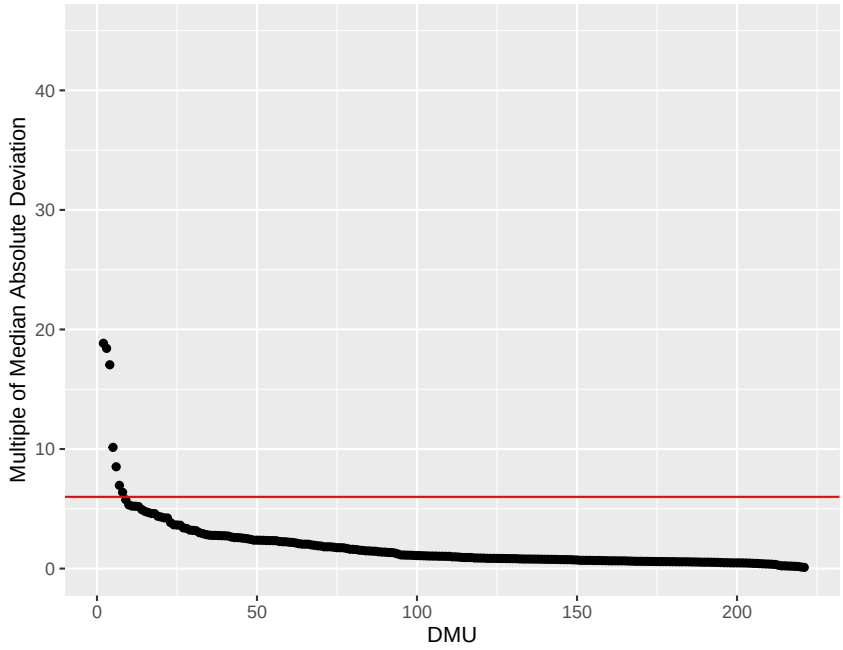


Figure 10: Visual representation of each DMU's distance to the median as a multiple of the median absolute deviation. The red line indicates six median absolute deviations from the median.

Further indication can be seen from the MAD distance graph presented in Figure 10. When the graph above is compared to Figure 5, it is clear that the observed points deviate less from the median value. The results of the efficiency scores for each DMU is, nonetheless, of more importance when determining whether the PCA has succeeded or not. The summary statistics of the computed efficiency scores are presented in Table 9.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.4148	0.7221	0.8259	0.7840	0.8854	1.0000

Table 9: Summary statistics of the computed efficiency scores.

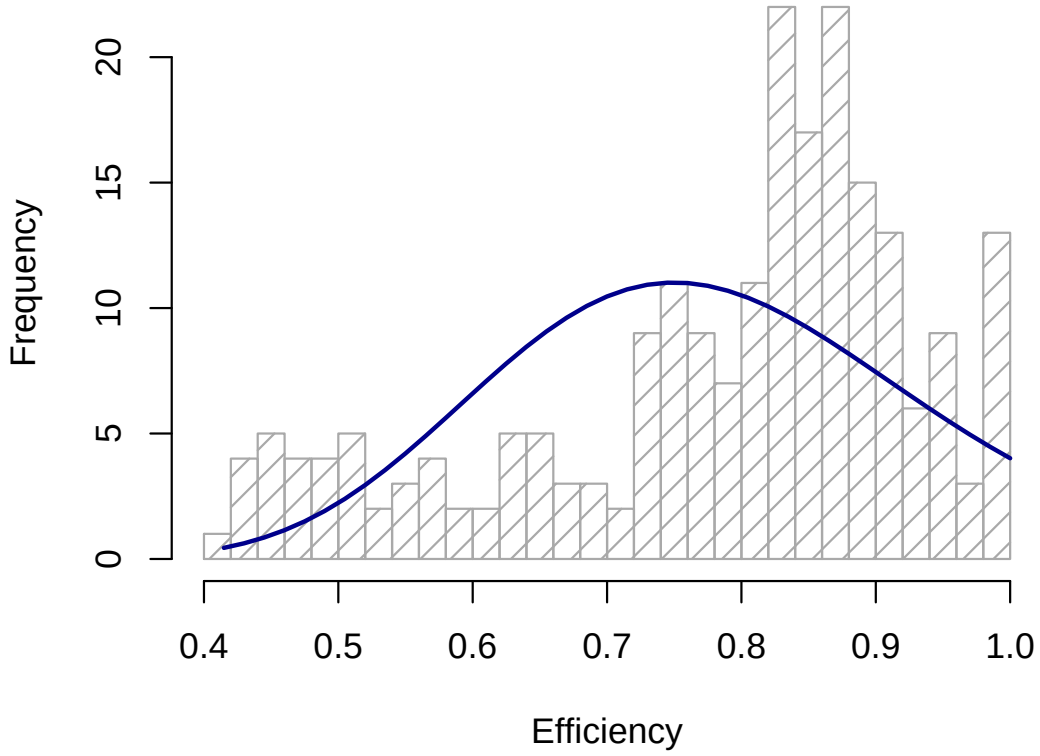


Figure 11: A fitted gamma curve ($\alpha = 23.13$, $\beta = 29.50$) over the frequency distribution of computed efficiency scores.

The resulting histogram of efficiency scores is presented in Figure 11. By comparing Figure 11 to Figure 6, it is clear that the DEA results have not been significantly affected by the non-centered data. The number of efficient units in this part was 7 compared to the 8 efficient units found with centered data. In addition, the histograms of efficiency scores are almost equally skewed.

3.5 Non-centered RPCA-DEA

In this part we perform the same computations as earlier using RPCA, but without centering the data in advance. When performing RPCA on the original input and output variables separately without centering the data, the resulting principal components are presented in Table 10.

	Input Variables			Output Variables	
	IPC1	IPC2	IPC3	OPC1	OPC2
Standard Deviation	1.1711	0.9564	1.0000	1.3953	1.0000
Proportion of Variance	0.7848	0.1223	−0.3967	0.9779	−0.4041
Cumulative Proportion	0.7848	0.9071	0.5104	0.9779	0.5738

Table 10: Summary of the principal components computed with RPCA without centering of data.

The share of variance explained by the first principal component has increased from 53.4% to roughly 78.5%, which is an almost identical increase as the one observed with PCA and no centering of data. In addition, the second principal component which accounted for a decrease in explained variance when data was centered now explains roughly 12.2% of the variance in the data, whereas the third principal component now obtained a negative value of approximately -40.4%.

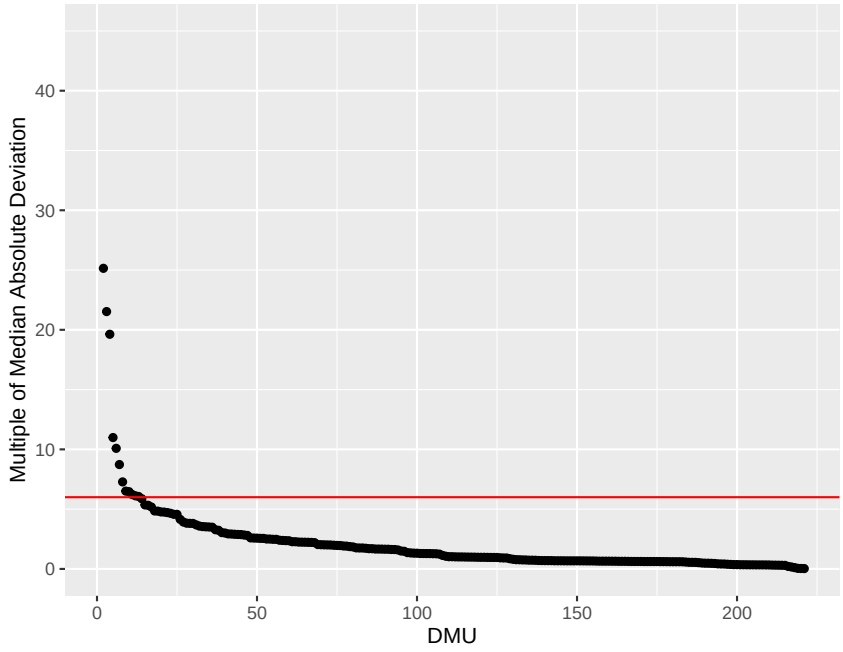


Figure 12: Visual representation of each DMU's distance to the median as a multiple of the median absolute deviation. The red line indicates six median absolute deviations from the median.

From Figure 12 it can be seen that the outlier values are deviating relatively more from the median value without centering of data, when comparing to Figure 7. Nonetheless, the cumulative variance explained by the first two principal components reaches the 80-90% interval described by Adler and Golany (2001), whereas the two main principal components computed with centered data only explain a total of 44.7% of the variance. The corresponding efficiency scores are presented in Figure 13.

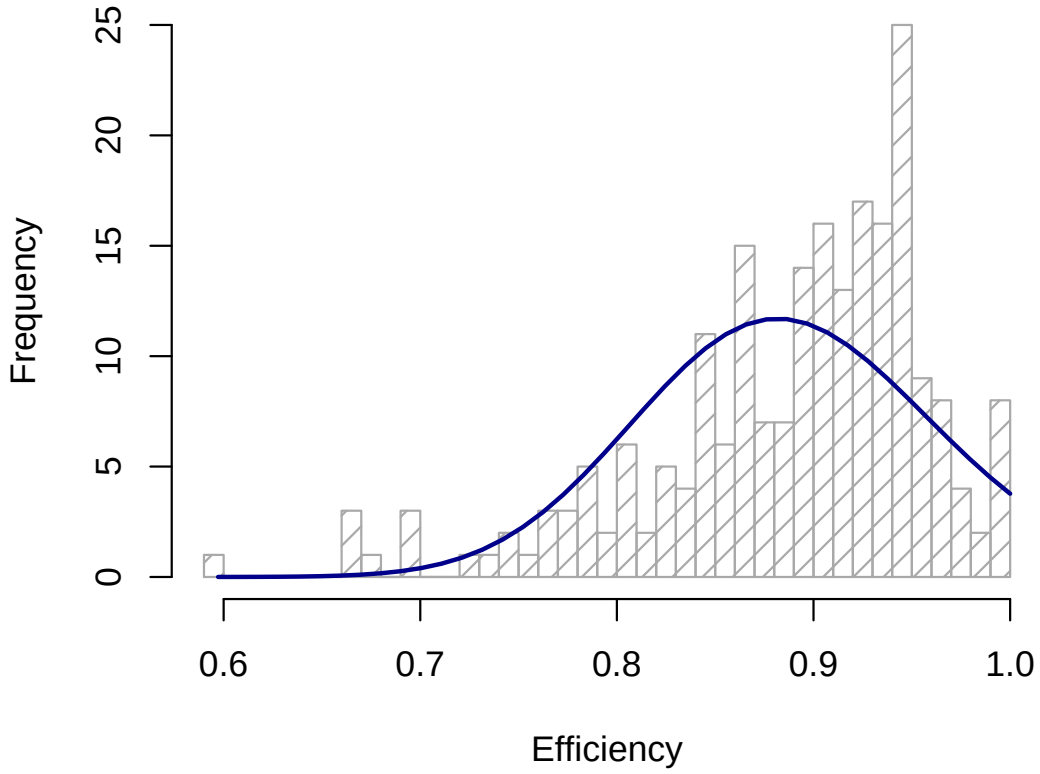


Figure 13: A fitted gamma curve ($\alpha = 138.11$, $\beta = 155.49$) over the frequency distribution of computed efficiency scores.

The summary statistics of the efficiency scores are presented in Table 11. The number of efficient units identified with RPCA and no centering of data was 7.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.5972	0.8538	0.9050	0.8882	0.9401	1.0000

Table 11: Summary statistics of the computed efficiency scores.

3.6 Non-scaled PCA-DEA

Performing PCA on the input and output variables without scaling data prior to the analysis yields results as presented in Table 12.

	Input Variables			Output Variables	
	IPC1	IPC2	IPC3	OPC1	OPC2
Standard Deviation	2224	3.887	0.727	762.957	26.304
Proportion of Variance	1.000	0.000	0.000	0.999	0.001
Cumulative Proportion	1.000	1.000	1.000	0.999	1.000

Table 12: Summary of the principal components computed with PCA without scaling of data.

The effect of not scaling the data prior to PCA is apparent, as indicated by the fact that the first principal component (IPC1) explains all of the variance of the input variables. In addition, the share of variance explained by the first principal component of the output variables (OPC1) has also increased to roughly 100%.

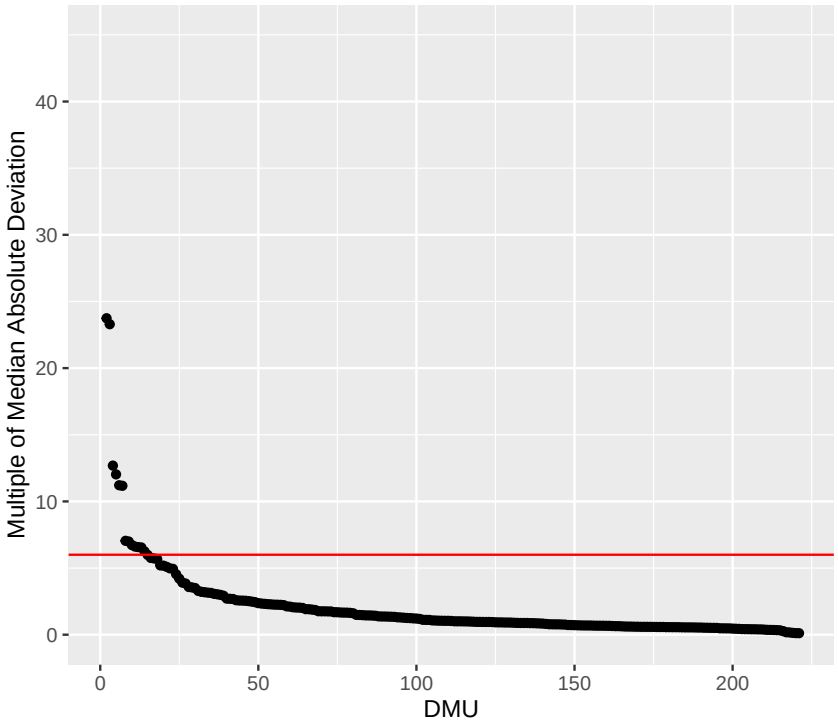


Figure 14: Visual representation of each DMU's distance to the median as a multiple of the median absolute deviation. The red line indicates six median absolute deviations from the median.

From Figure 14 we can notice that the outlier values are forming clear clusters relative to the majority of the data points. When comparing the distance plot to Figure 5 and 10, it is clear that the outliers are still significantly present, although less so than when both centering and scaling of data was performed before PCA. The corresponding histogram of efficiency scores from the DEA calculation are presented in Figure 15.

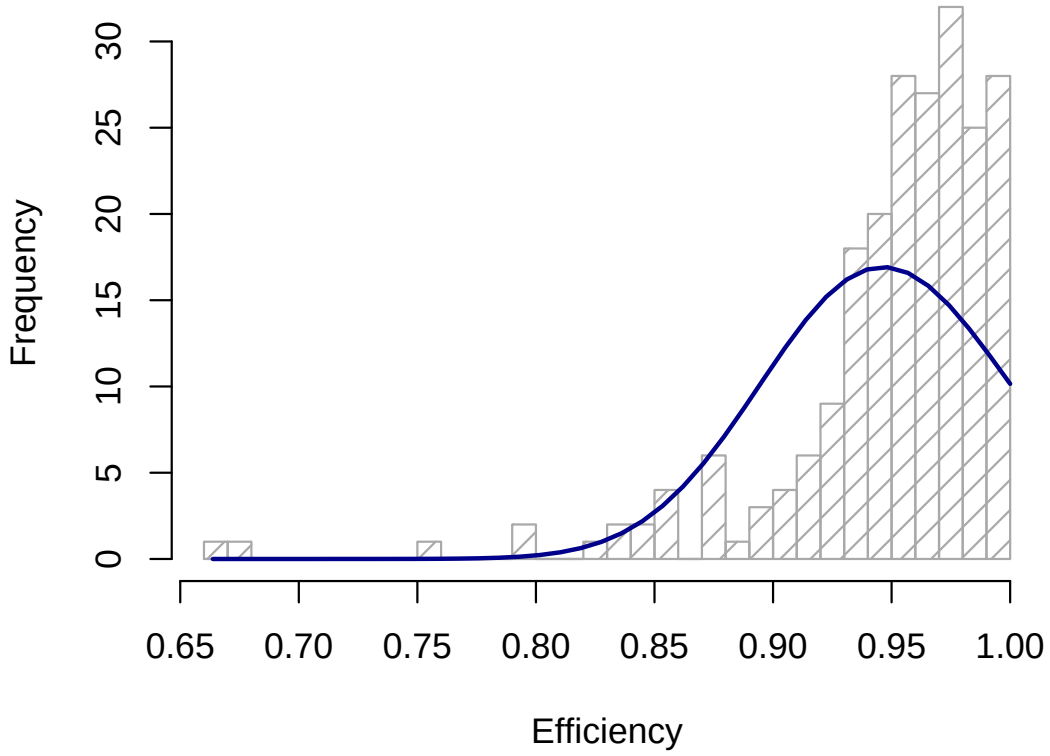


Figure 15: A fitted gamma curve ($\alpha = 331.18$, $\beta = 348.89$) over a frequency distribution of computed efficiency scores.

The summary statistics of the efficiency scores are presented in Table 13. The number of efficient units identified with RPCA and no centering of data was *13*.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.6637	0.9371	0.9609	0.9492	0.9787	1.0000

Table 13: Summary statistics of the computed efficiency scores.

3.7 Non-scaled RPCA-DEA

The resulting principal components from performing RPCA without scaling of data are presented in Table 14.

	Input Variables			Output Variables	
	IPC1	IPC2	IPC3	OPC1	OPC2
Standard Deviation	2224.27	4.470	0.750	762.951	91.619
Proportion of Variance	1.000	0.000	0.000	0.999	−0.013
Cumulative Proportion	1.000	1.000	1.000	0.999	0.986

Table 14: Summary of the principal components computed with RPCA without scaling of data.

The same effect can be witnessed here as with regular PCA when the data is not scaled. The first principal components are solely explaining all of the variance in the data. However, the relative deviations between the data points are not as clear with RPCA as with PCA, which is apparent when comparing Figure 14 to Figure 16 below.

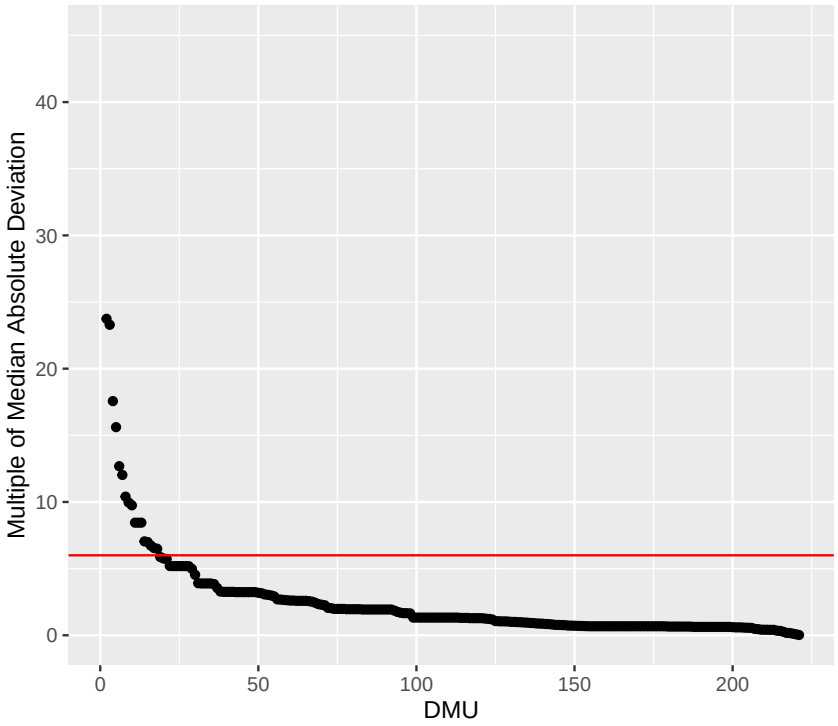


Figure 16: Visual representation of each DMU's distance to the median as a multiple of the median absolute deviation. The red line indicates six median absolute deviations from the median.

Figure 16 shows clearly that the outlier values are deviating more from the median when data has not been scaled, as compared to when data is both scaled and centered. This can be visually confirmed by comparing Figure 7 to the graph above. A histogram of the efficiency scores from the subsequent DEA is presented in Figure 17.

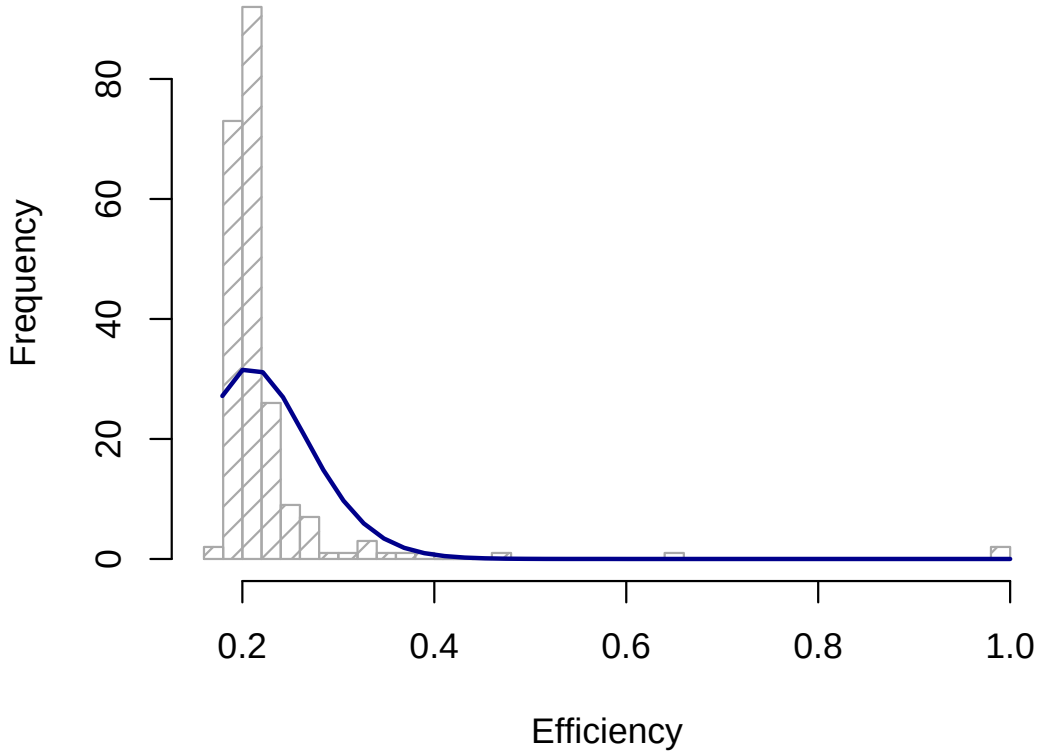


Figure 17: A fitted gamma curve ($\alpha = 15.44$, $\beta = 69.17$) over the frequency distribution of computed efficiency scores.

The summary statistics of the efficiency scores are presented in Table 15. The number of efficient units identified with RPCA and no scaling of data was *2*.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.1792	0.1950	0.2023	0.2233	0.2198	1.0000

Table 15: Summary statistics of the computed efficiency scores.

3.8 Non-scaled, non-centered PCA-DEA

In the last two parts we perform the same computations as in the previous stages, but with the difference that the data used for PCA and RPCA, respectively, has neither been centered nor scaled. Using the aforementioned setup and regular PCA yields the principal components presented in Table 16.

	Input Variables			Output Variables	
	IPC1	IPC2	IPC3	OPC1	OPC2
Standard Deviation	2442	5.929	0.945	833.139	26.797
Proportion of Variance	1.000	0.000	0.000	0.999	0.001
Cumulative Proportion	1.000	0.000	0.000	0.999	1.000

Table 16: Summary of the principal components computed with PCA without centering or scaling of data.

It is clear that the effects of not scaling the data are present, as the variances explained by each principal component are virtually equal to the results obtained with PCA and no scaling of data. This can be visually confirmed by comparing the table above to Tables 12 and 14.

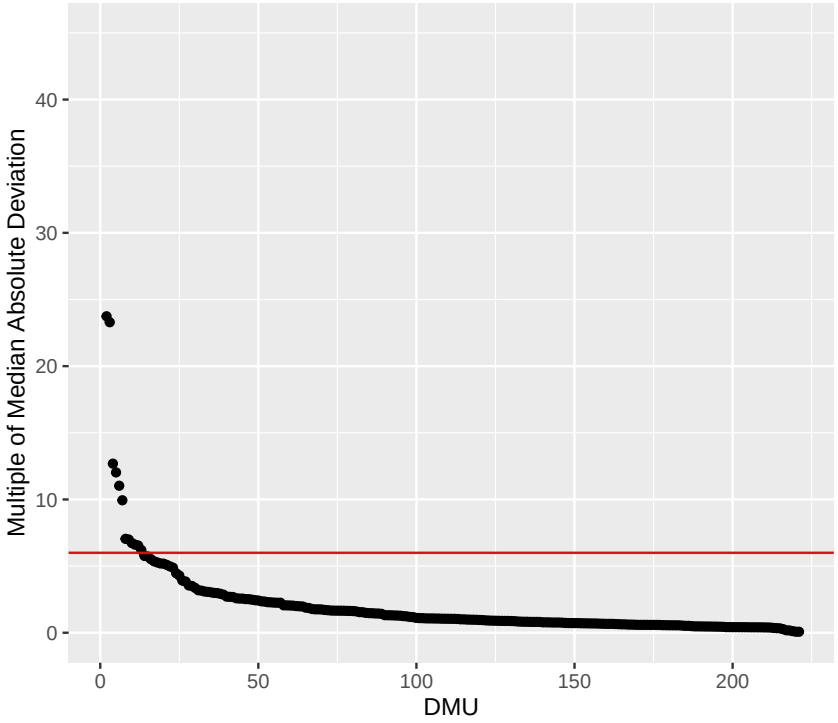


Figure 18: Visual representation of each DMU's distance to the median as a multiple of the median absolute deviation. The red line indicates six median absolute deviations from the median.

It can be seen from Figure 18 that the outlier values deviate more from the median value when the data is not scaled prior to PCA, as can be confirmed by comparing the graph above to Figures 10 and 14. A histogram of the computed efficiency scores is presented in Figure 19.

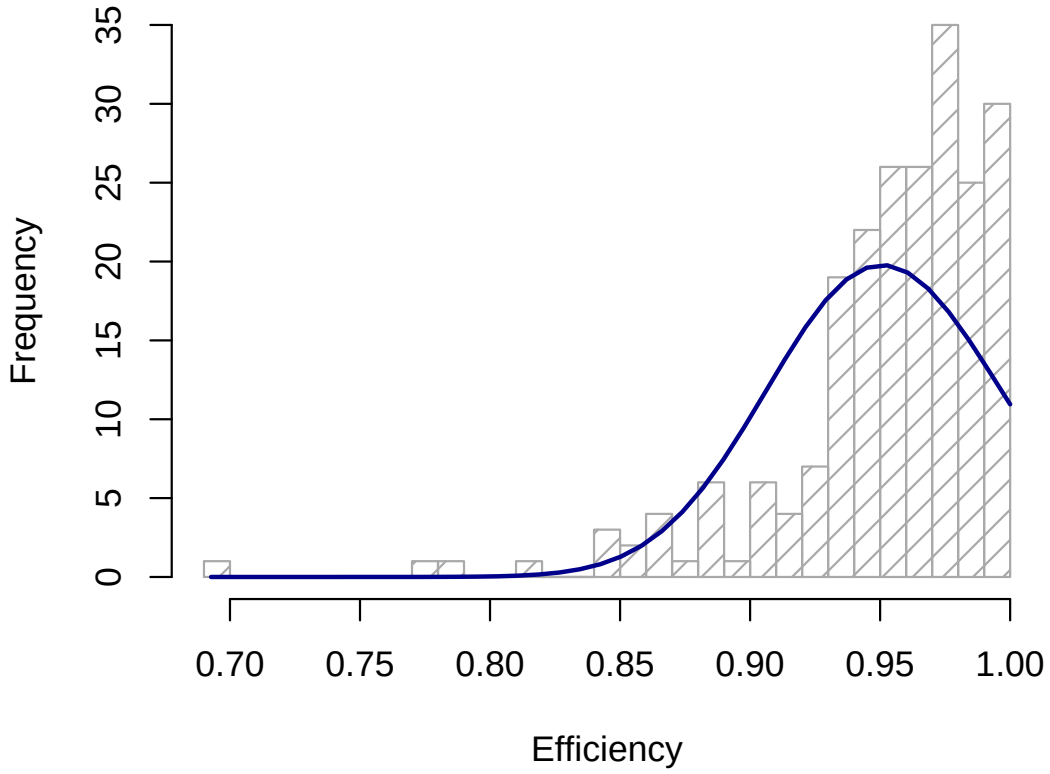


Figure 19: A fitted gamma curve ($\alpha = 456.10$, $\beta = 478.70$) over the frequency distribution of computed efficiency scores.

The summary statistics of the efficiency scores are presented in Table 17. The number of efficient units identified with PCA without centering or scaling of data was *11*.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.6927	0.9396	0.9625	0.9528	0.9799	1.0000

Table 17: Summary statistics of the computed efficiency scores.

3.9 Non-scaled, non-centered RPCA-DEA

In this last part we perform RPCA on the input and output variables separately without centering or scaling the data. The obtained principal components are then used in the subsequent DEA model. The principal components with the aforementioned setup are presented in Table 18.

	Input Variables			Output Variables	
	IPC1	IPC2	IPC3	OPC1	OPC2
Standard Deviation	2224.265	4.418	0.750	762.929	91.619
Proportion of Variance	1.000	0.000	0.000	0.999	−0.014
Cumulative Proportion	1.000	0.000	0.000	0.999	0.985

Table 18: Summary of the principal components computed with RPCA without centering or scaling of data.

The same issue is apparent here as with the previous combinations of not scaling the data before computing principal components. IPC1 and OPC1 are measured in a completely different magnitude than the other variables, naturally resulting in the two variables accounting for virtually all variance in the data.

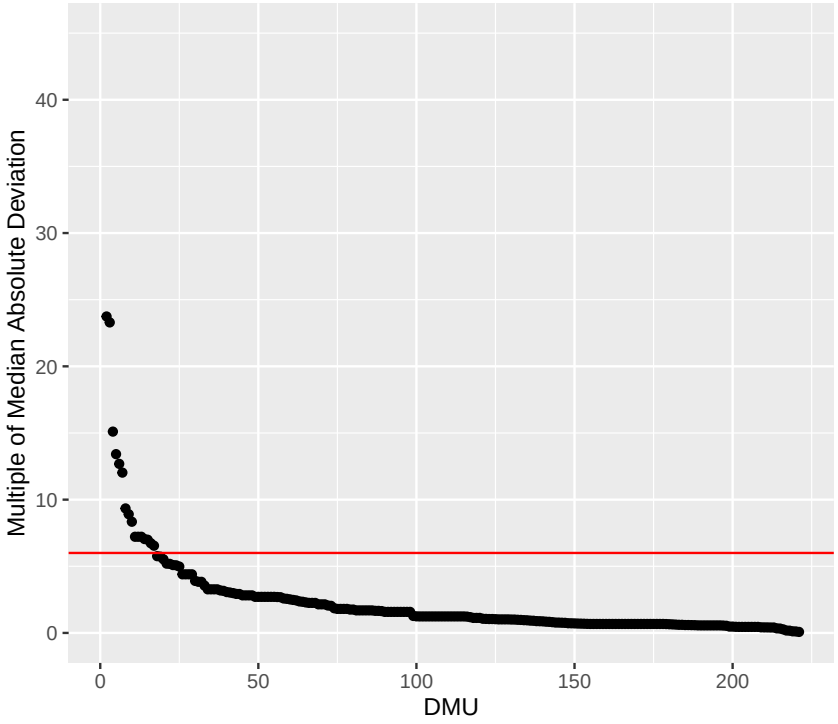


Figure 20: Visual representation of each DMU's distance to the median as a multiple of the median absolute deviation. The red line indicates six median absolute deviations from the median.

No significant improvement or deterioration can be seen when comparing Figure 20 to the three other distance plots from the setups where scaling was not performed. The corresponding histogram of efficiency scores is presented in Figure 21.

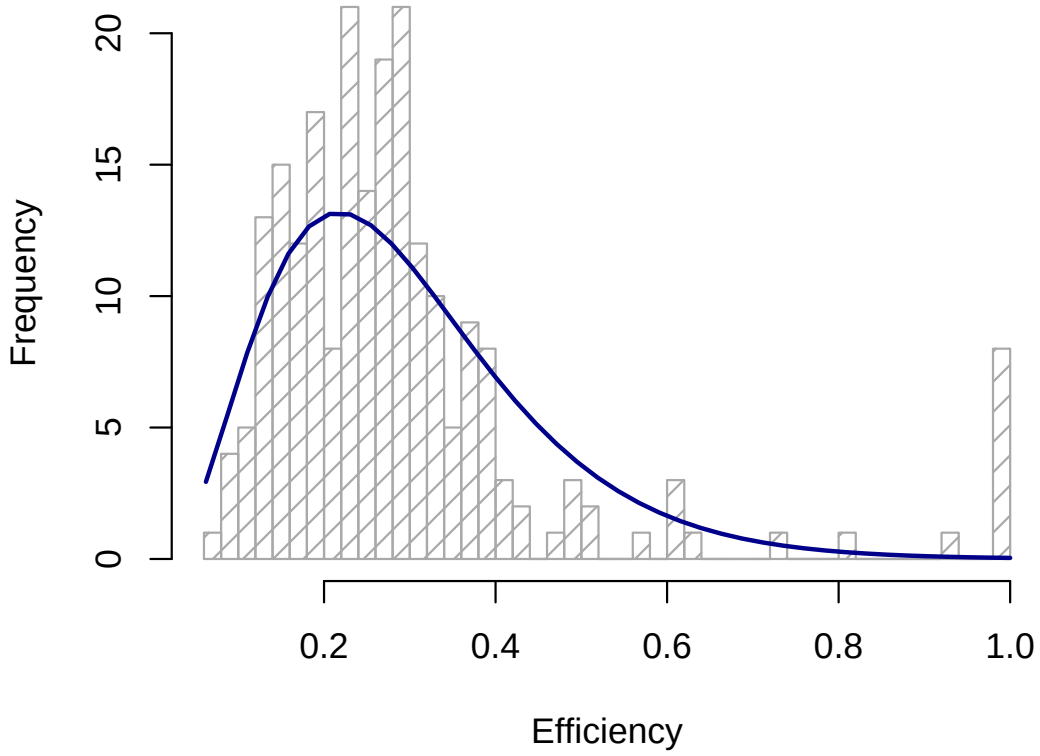


Figure 21: A fitted gamma curve ($\alpha = 3.80$, $\beta = 12.88$) over the frequency distribution of computed efficiency scores.

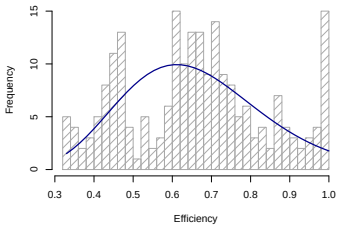
The summary statistics of the efficiency scores are presented in Table 19. The number of efficient units identified with RPCA without centering or scaling of data was *8*.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0623	0.1899	0.2602	0.2950	0.3274	1.0000

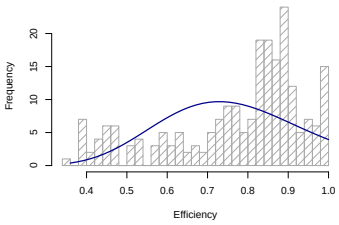
Table 19: Summary statistics of the computed efficiency scores.

3.10 Compilation of results

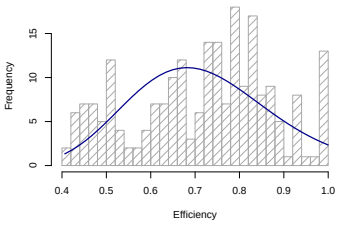
The results presented in this section have been compiled in this subsection to summarize the most clear discoveries and to allow the results to be compared with one another. All histograms of efficiency scores have been compiled in Figure 22 on the next page.



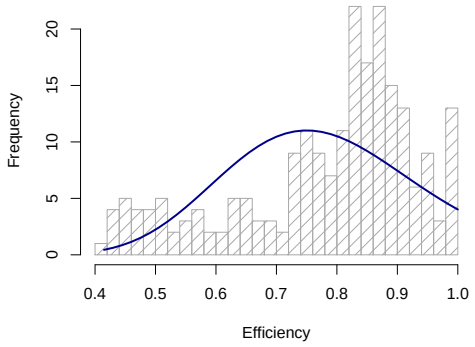
(a) DEA



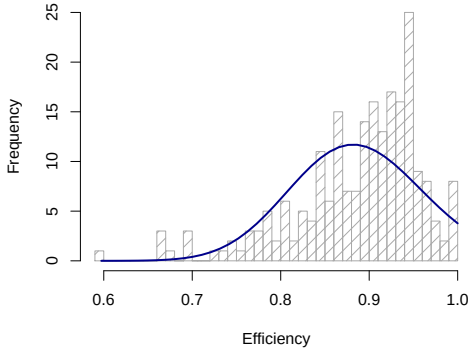
(b) PCA-DEA



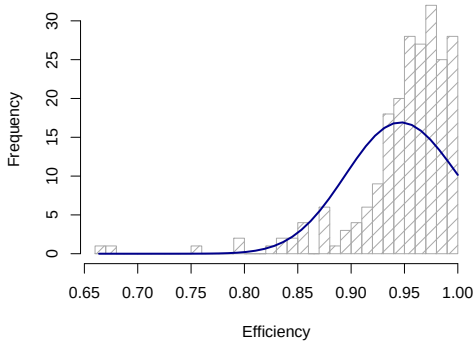
(c) RPCA-DEA



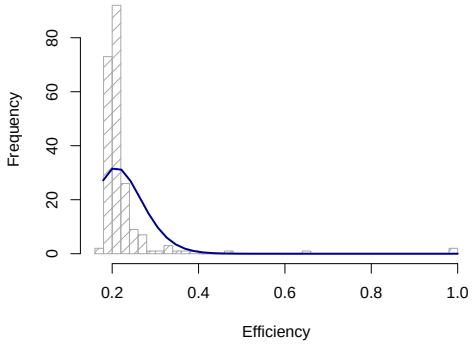
(d) PCA-DEA not centered



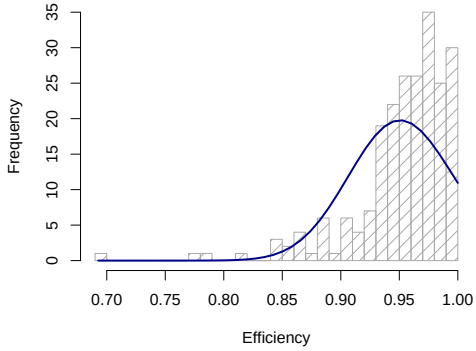
(e) RPCA-DEA not centered



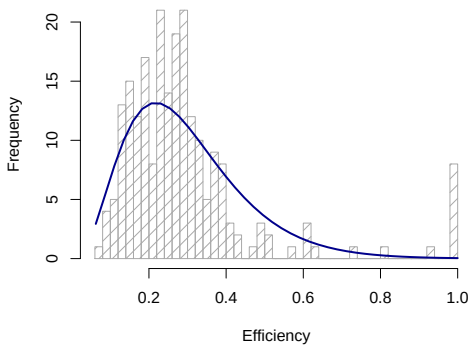
(f) PCA-DEA not scaled



(g) RPCA-DEA not scaled



(h) PCA-DEA not centered or scaled



(i) RPCA-DEA not centered or scaled

Figure 22: Compilation of all histograms of efficiency scores.

By looking at Figures 22b and 22d and comparing them to Figures 22f and 22h, it is clear that scaling causes the computed efficiency scores to become more negatively skewed with PCA-DEA. On the other hand, with RPCA-DEA scaling seems to cause the opposite effect, as the efficiency scores became more positively skewed as can be seen from Figures 22g and 22i. From these results we can conclude that scaling should always be performed prior to variable reduction techniques. However, no clear conclusions could be drawn by only inspecting the efficiency scores regarding whether centered data or non-centered data provides more reliable results. The advantage of using a more robust estimator is also not clear by only inspecting the computed efficiency scores, as Figures 22b and 22d are fairly similar to Figures 22c and 22e.

4 Conclusions

The purpose of this thesis was to perform a case study on Scandinavian retail companies using *Data Envelopment Analysis* and two variations of *Principal Component Analysis* in order to study how the results of DEA are affected by the usage of variable reduction. The results of DEA are of particular importance in data-driven decision making, and as data is increasingly gathered and utilized, so is the need for more precise results. In addition, decision making is dependent on an increasing number of variables, which raises the need for a secure way of dimensionality reduction without loss of information.

The study was performed initially in three stages, where the baseline for the DEA results was first set by performing DEA on the entire data set using all three input variables and both output variables. In the second part, PCA was first conducted on the input and output variables separately to obtain principal components. The first two principal components of the input variables were used as input in the subsequent DEA model, whereas the first principal component of the output variables was the sole output. In the third stage, *Robust PCA* was used to obtain principal components which have not been affected by outliers and other corrupted observations as much as when regular PCA is used. The input and output variables in the subsequent DEA model were otherwise equal.

As both PCA and DEA are sensitive to outliers, we wanted to investigate how *centering* and *scaling* of data prior to the dimensionality reduction affects the scores from DEA. For this purpose, both PCA-DEA and RPCA-DEA calculations were repeated a total of three times, first by not centering the data, then without scaling it, and lastly without centering or scaling the data.

The results from initial three stages where data was both scaled and centered indicate that outliers have significantly affected the principal components used in the DEA model. This can be confirmed from the results of the PCA, and from the fact that the set of efficiency scores is more negatively skewed than when DEA is performed without any form of variable reduction. However, the presence of outliers might have

improved the overall results of DEA as has been argued in previous research. The fact that the number of efficient units decreased in both PCA-DEA and RPCA-DEA from the initial 13 units in conventional DEA supports this claim. On the other hand, this decrease in efficient units could also be caused by the decrease from in number of variables used in computing the efficiency scores.

By repeating the PCA-DEA and RPCA-DEA calculations using the three combinations of centering and scaling, we obtained results that clearly support the usage of scaling the data prior to dimensionality reduction. The resulting principal components are heavily affected by differences in scale between the variables, in this case leading to one variable accounting for virtually all variance in the data. A clear difference in efficiency results is shown between PCA-DEA and RPCA-DEA, however, as scaling of the data shifts the distribution of efficiency scores towards the maximum value in PCA-DEA. On the other hand, scaling of data shifts the distribution of scores to the opposite direction when RPCA-DEA is used, leading to a positively skewed distribution. The effects of centering the data are not as clear, as the distribution of efficiency scores remained almost equal in PCA-DEA, and the number of efficient units decreased by one. There were more apparent effects of centering in RPCA-DEA, as the distribution of efficiency scores was more negatively skewed than when the data was centered. Furthermore, the number of efficient units decreased by two.

The biggest improvement was, however, in the obtained principal components from RPCA. When the data was centered the first two PCs accounted for less than 50% of the variance in the data, whereas the same PCs accounted for more than 90% of the total variance when data was not centered. In combination with the results from the subsequent DEA calculation, the non-centered RPCA-DEA model presents possibly the most confident results in this study, rivaling those of the regular RPCA-DEA model.

5 Future prospects

There is still room for further research on the effects of PCA and RPCA on a subsequent DEA calculation. One of the main points of interest could be in the investigation of how the computed efficiency scores are affected if all outliers are removed from the data set. In this study, we chose not to remove the outliers, although they were significantly affecting both the PCA and the DEA calculations.

Another important aspect of further research could be to investigate when it is necessary to use variable reduction techniques before a DEA calculation. In this study, the number of variables was relatively low (3 inputs, 2 outputs), which means that there was no restricting factor that forced us to perform variable reduction. This raises the question whether the variable reduction actually improved the results or not.

A third prospective field of research considers the assumption of returns to scale in the DEA model. We used the BCC model in our calculations, meaning that we assumed variable returns to scale. However, there is no evidence that this assumption is correct. Further studies could therefore focus on the differences in results when the assumption of returns to scale is changed between increasing, decreasing, constant, and variable returns to scale.

References

Nicole Adler and Boaz Golany. Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to Western Europe. *European Journal of Operational Research*, 132(2): 260–273, July 2001. ISSN 0377-2217. doi: 10.1016/S0377-2217(00)00150-8. URL http://www.sciencedirect.com/science/article/pii/S0377221700001508.

Nicole Adler and Ekaterina Yazhemsky. Improving discrimination in data envelopment analysis: PCA–DEA or variable reduction. *European Journal of Operational Research*, 202(1):273–284, April 2010. ISSN 0377-2217. doi: 10.1016/j.ejor.2009.03.050. URL http://www.sciencedirect.com/science/article/pii/S037722170900229X.

R. D. Banker, A. Charnes, and W. W. Cooper. Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science*, 30 (9):1078–1092, September 1984. ISSN 0025-1909. doi: 10.1287/mnsc.30.9.1078. URL https://pubsonline.informs.org/doi/abs/10.1287/mnsc.30.9.1078.

A. Charnes, W. W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444, November 1978. ISSN 0377-2217. doi: 10.1016/0377-2217(78)90138-8. URL http://www.sciencedirect.com/science/article/pii/0377221778901388.

Naveen Donthu and Boonghee Yoo. Retail productivity assessment using data envelopment analysis. *Journal of Retailing*, 74(1):89–105, March 1998. ISSN 0022-4359. doi: 10.1016/S0022-4359(99)80089-X. URL http://www.sciencedirect.com/science/article/pii/S002243599980089X.

Naveen Donthu, Edmund K. Hershberger, and Talai Osmonbekov. Benchmarking marketing productivity using data envelopment analysis. *Journal of Business Research*, 58(11):1474–1482, November 2005. ISSN 0148-2963. doi: 10.1016/j.jbusres.2004.05.007. URL http://www.sciencedirect.com/science/article/pii/S0148296304001778.

SC Hui and MC Wan. Study of hotel energy performance using data envelopment analysis. 2013. URL https://www.researchgate.net/publication/281901553_Study_of_hotel_energy_performance_using_data_envelopment_analysis.

Mohamed M. Mostafa. Benchmarking the US specialty retailers and food consumer stores using data envelopment analysis. *International Journal of Retail & Distribution Management*, 37(8):661–679, January 2009. ISSN 0959-0552. doi: 10.1108/09590550910966178. URL https://doi.org/10.1108/09590550910966178.

Reet Põldaru and Jüri Roots. A PCA–DEA approach to measure the quality of life in Estonian counties. *Socio-Economic Planning Sciences*, 48(1):65–73, March 2014. ISSN 0038-0121. doi: 10.1016/j.seps.2013.10.001. URL http://www.sciencedirect.com/science/article/pii/S0038012113000517.

Lawrence M. Seiford. A bibliography for Data Envelopment Analysis (1978-1996). *Annals of Operations Research*, 73(0):393–438, October 1997. ISSN 1572-9338. doi: 10.1023/A:1018949800069. URL https://doi.org/10.1023/A:1018949800069.