Erasmus University Rotterdam
Erasmus School of Economics

Master Thesis Business Analytics and Quantitative Marketing

A machine Learning Approach to Extreme Value Analysis

Name student: Dennis Ruimschoot
Student ID Number: 476933

Supervisor: dr. Phyllis Wan
Second assessor: Jochem Oorschot

Date final version: 3rd July 2021

A machine Learning Approach to Extreme Value Analysis

Dennis Ruimschoot

3rd July 2021

Abstract

Over the last years research on extreme values has become increasingly popular. Observations are classified as extreme values if they are in the heavy tails. We look at three methods in Extreme Value Analysis: spherical k-means, Principal Component Analysis (PCA) and spherical k-principal-components. We perform all methods on a finance data set and two pollution data sets. Spherical k-means and PCA show that extreme losses in heavy industries coincide with each other, while being mitigated by losses in consumer products. Spherical k-principal-components shows a similar result when using random starting values. All methods indicate that pollution of any type is mostly observed independently of other types. Spherical k-principal-components breaks down in high dimensions when using k-means starting values. The main finding of this thesis is that all three methods produce almost identical results for low dimensional data, while producing some overlapping results for high dimensional data.

Contents

1 Introduction

Finding patterns in data is something that researchers have been doing for a long time. Researches have come up with various methods for different types of data to try and get an idea of the properties of the data. One type of analysis that has become very popular over the last couple of years is called Extreme Value Analysis. As the name suggests, Extreme Value Analysis deals with analysing observations that are in the extreme parts of the tail belonging to a distribution. The main reason for this spike in interest in Extreme Value Analysis is the ongoing climate change. Extreme temperatures as well as other extreme scenarios like floodings and hurricanes are not as rare as they used to be, which makes it easier to investigate these occurrences (Easterling et al., 2000). Moreover, Extreme Value Analysis has not only been applied to extreme weather conditions, but also to financial applications for example in which someone can assess the risk of a client not paying its loan (Brodin & Klüppelberg, 2014).

This thesis looks at three different Extreme Value Analysis methods performed on finance and pollution data to answer the research question: "Which clustering methods/dimension reduction algorithms can we combine with Extreme Value Analysis?" The first method uses Extreme Value Analysis in conjunction with spherical k-means, a method similar to "standard" k-means but with the main difference being that spherical k-means looks at an angular measure for the distance instead of the Euclidean distance. The second method makes use of a special form of Principal Component Analysis (PCA) to see which variables are most important in explaining the variability that is present in the data. The final method we use is called spherical k-principal-components and, as the name suggests, combines the two previously explained methods. Section 2 discusses previously conducted research on similar topics as well as knowledge that is required to understand the subject. Section 3 discusses that data used to answer the research question and the transformations performed on the raw data. Section 4 goes into depth on the mathematics behind the use of spherical k-means, PCA and spherical k-principal-components for Extreme Value Analysis. Section 5 shows the main results that are obtained from performing all methods. Section 6 concludes with a conclusion on the obtained results as well as a short discussion with some future recommendations. This thesis combines three unsupervised learning methods with Extreme Value Analysis, so this thesis adds to the current Extreme Value Analysis literature as well as to the existing machine learning literature. The main contribution of this thesis to the present literature is that we use all methods for different data sets, so that we highlight situations in which any of the methods do not perform well. In addition we mention some limitations. The Conclusion section also explains how we can relate the results from the three different methods so that we can accurately compare their outcomes, which has not been done so far.

2 Literature

This section touches on previously conducted research on similar topics as well as introducing vital concepts that are important to understand the methodology.

2.1 k-means

Data that has to be analysed is often very large. For example one could analyse data of all households in the US for 20 years or look at monthly precipitation data in the country of Portugal. These data sets would have many observations which is beneficial, but the added downside of this fact is that for both data sets we cannot assume that all data was observed under the same conditions. In case of the household data different individuals could for example have come from a different city. For the precipitation data it is unreasonable to assume that in the month of January we observe as much rain as in the month of July. To account for these differences among various observations we try to incorporate so-called heterogeneity. Sometimes it is also referred to as unobserved heterogeneity, because the differences may not be observable (Mannering et al., 2016). By not accounting for individual differences we would assume all data to have the same characteristics, which in turn would lower the accuracy of our results. A regularly implemented approach to incorporate heterogeneity is called k-means. As the name of this method somewhat implies, the goal is to investigate the means of different groups of data. Beforehand we specify the number of groups k, also referred to as clusters, in which we segment the data. The goal is then to assign observations with similar characteristics in one cluster, while putting observations with distinct characteristics in different clusters (Likas et al., 2003). The name k-means could make one think that we can only investigate the means, but this is not the case. While we do assign observations optimally based on their "distance" from the mean, each cluster is still a subset of the data. This means that we can perform any known method of analysing data on the observations in a cluster. Note that all variables should be of the same scale, otherwise the optimisation algorithm mainly focuses on variables with large variance. One limitation of this method however is that (assuming all variables are normalised) the method assumes all variables to be of equal importance. In other words, it assumes that every single variable contributes equally to the variation of observations. A method that specifically looks at the importance of each variable is called Principal Component Analysis.

2.2 Principal Component Analysis

Not all variables in a data set contribute as much to the structure of the used data. Pearson (1901) therefore introduced a method to explain the variance-covariance structure of the variables which he called Principal Component Analysis (PCA). The idea behind PCA is that we construct new variables called (principal) "components" that are linear combinations of all the original variables. We obtain all coefficients by calculating the eigenvectors for the covariance matrix. Then by looking at the coefficient for each variable, we can see how heavily the component depends on each variable

(Drees & Sabourin, 2019). Next we obtain the eigenvalues of the same covariance matrix. Because the covariance matrix is always positive semi-definite by construction, all eigenvalues are non-negative and we can make a ranking of all eigenvalues. Each component has one corresponding eigenvalue, so the component with the largest eigenvalue has the largest variance and is thus most influential in explaining the variability of the data. The component corresponding to the second largest eigenvalue then has the second largest variance, and so on until we get to the component with the smallest eigenvalue and hence the smallest variance (Johnson & Wichern, 2014). For example we could look at a data set containing the variables rainfall, sunshine and temperature. If the component with the largest eigenvalue has a high coefficient for sunshine and temperature, but not for rainfall, then sunshine and temperature are more important in explaining the variance of the data than rainfall. Sometimes the new components have a logical interpretation based on the coefficients. A medical example could be that a component has large coefficients for the variables diabetes and blood pressure. The component now signifies the likelihood of having heart failure as having diabetes and a high blood pressure both increase the change of having a heart failure (From et al., 2006). Be aware that when one wants to use PCA, the data should first be normalised so that every variable has the same scale. If the researcher does not do this beforehand, PCA will focus more heavily on variables with a larger variance and the conclusion will always be that the variables with a high variance are most important in explaining the total variance of the data.

2.3 Extreme Value Analysis

Different types of data require different types of models to investigate the data. For example data recorded over time requires a time series model, while for cross-sectional data a simple linear model may suffice. A special type of analysis that deals with observations which are very far away from the bulk of the data, also referred to as observations in the heavy tails, is called Extreme Value Analysis. Tippett (1925) pioneered this special type of unsupervised learning, but most research on this topic has been performed in the last few decades due to climate change. Extreme events such as high temperatures and flooding are becoming more common which makes it easier for researchers to investigate these occurrences. A concise summary of relevant papers on Extreme Value Analysis is given by Zhang et al. (2004). Extreme Value Analysis is a collective name that encompasses many different methods, but all methods have something in common. In Extreme Value Analysis we always start off by examining which observations in the data can be constituted as extreme values. Usually we first apply a form of normalisation to ensure that all variables lie within a common interval, after which the 5-10% largest observations become the new data set. Now we perform any known method to analyse the extreme values. Often we can perform conventional methods on the extreme values, for example the Bayesian mixture model by Kottas & Sansó (2007). However for some methods we need to take extra steps to account for the fact that we are dealing with extreme values. This is for example the case when trying to perform Principal Component Analysis as proposed by Cooley & Thibaud (2019). In this paper Cooley & Thibaud use an altered formula

to estimate the covariance matrix. This formula is tweaked so that it can be estimated for extreme values. The "regular" unbiased estimator for the covariance matrix is not used.

3 Data

This section discusses all data used to answer the research question. We use two different data sets. The first data set is referred to as "*finance*" and contains data on daily stocks of companies specialised in various markets. Data is observed from July 1st 1926 up to March 31st 2021. This data set is obtained from the Kenneth French Data Library[1]. Among other things, the data set contains data for companies dealing with food, beer, smoking, oil, etc. The data set also contains one variable that indicates on which day the stock was recorded. This variable is only used to filter the data, as we would only like to model stocks observed during the time period 1950-2015. This variable is subsequently removed, which leaves us with $p = 30$ variables. The number of observations after filtering is $n = 16,694$. Companies can of course record either losses or profit, which is why the data set contains both positive and negative values. For this research however we are only interested in extreme losses, i.e. the extreme negative values. Therefore we transform the observations using the formula

$$x_{ij}^{(temp)} = max\{-x_{ij}, 0\},\ i = 1, \ldots, n,\ j = 1, \ldots, p. \tag{1}$$

This transformations ensures that all gains are zero, while losses are now larger than zero. Finally we apply a transformation to this new temporary data set to ensure that those values are bounded away from zero. The formula for this transformation is the following:

$$\hat{x}_{ij} = t(x_{ij}^{(temp)}) = ln[1 + exp(x_{ij}^{(temp)})],\ i = 1, \ldots, n,\ j = 1, \ldots, p. \tag{2}$$

The expression $ln(\cdot)$ in Equation (2) signifies the natural logarithm. The transformation in Equation (2) only has effect on small values, as the value 1 within the logarithm is negligible if $x_{ij}^{(temp)}$. This is the case because for large $x_{ij}^{(temp)}$ the term $exp(x_{ij}^{(temp)})$ becomes very large and so $x \approx ln[1+exp(x)]$ (Cooley & Thibaud, 2019). We will perform our methods using the finance data set for the data points $\hat{x}_{ij}$.

The second data set comes from the R package **texmex**[2]. This data set is referred to as *pollution* and contains data on several air pollution statistics measured over the 1994-1998 period in the British city of Leeds. This data set is comprised of 5 variables; daily maximum ozone (O_3) measured in parts per billion, daily maximum nitrogen dioxide (NO_2) in parts per billion, daily maximum nitric oxide (NO) in parts per billion, daily maximum sulfur monoxide (SO_2) in parts per billion and particulates (PM_{10}) measured in micrograms/metre^3. The data is segmented into winter months (November up to and including February) and summer months (April up to and including July), which can be

[1] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
[2] https://cran.r-project.org/web/packages/texmex/texmex.pdf

obtained with the commands *data(winter)* and *data(summer)* respectively. The winter data set contains a total of 532 observations, while the summer data set contains a bit more observations with 578. Summary statistics for the winter months are shown in Table 1.

Table 1: Summary statistics winter air pollution

Statistic	O_3	NO_2	NO	SO_2	PM_{10}
Mean	20.056	44.169	135.457	21.032	48.442
Median	22	43	112.5	15	40
Maximum	44	104	568	200	177
Minimum	1	19	10	1	7
Std. Dev	10.896	11.287	102.487	20.951	27.894

The summary statistics for the summer months are shown in Table 2.

Table 2: Summary statistics summer air pollution

Statistic	O_3	NO_2	NO	SO_2	PM_{10}
Mean	31.997	37.628	55.201	17.369	41.123
Median	31	36	47	8	34
Maximum	84	105	256	313	185
Minimum	8	9	4	0	9
Std. Dev	10.335	11.862	36.443	29.264	23.133

The two data sets contain a different number of observations and the summary statistics indicate that the two data sets are not identically distributed. There is no feasible way of obtaining a one-to-one relationship between the observations in both data sets or combining the data in any other way, so we will examine the data set separately.

4 Methodology

This section discusses the main methods that are implemented. First we introduce the concept of extreme values and regular variation. After that we will discuss spherical k-means, Principal Component Analysis and spherical k-principal-components.

4.1 Generalised Extreme Value Distribution

This research deals with analysing extreme values in data. We would thus like to obtain a distribution for the maximum of a set of data points, where the maximum is denoted as $M_n = max\{x_1, ..., x_n\}$ and $\{x_1, ..., x_n\}$ is the sequence of independent identically distributed (iid) data points. As all data points are smaller than or equal to the data point with the maximum value, a naïve approach of obtaining a distribution for M_n is to calculate the probability of all data points lying below a certain threshold. Mathematically this is equal to

$$\begin{aligned} Pr[M_n \leq z] &= Pr[x_1 \leq z, ..., x_n \leq z] \\ &= Pr[x_1 \leq z] \times \cdots \times Pr[x_n \leq z] \\ &= [F(z)]^n. \end{aligned} \tag{3}$$

We can compute all n probabilities separately, because we assumed each data point to be independent. The main issue with the approach in Equation (3) however is that the distribution function $F(\cdot)$ is usually not known (Coles et al., 2001). Also, if the distribution function is different over individuals then we cannot use the final expression in (3), regardless of whether we know every distribution. On thing that we can use to obtain the distribution function is to use a linear renormalisation $M_n* = \frac{M_n - b_n}{a_n}$, $a_n > 0$, $b_n \in R$, so that we obtain a distribution function as introduced by Jenkinson (1955):

$$P(M_n* \leq z) \approx G(z) = exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi} \right\}, \; 1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0. \tag{4}$$

The corresponding probability density function is now

$$g(z) = G'(z) = \frac{\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi - 1}}{\sigma} exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi} \right\}. \tag{5}$$

The distribution associated with Equation (4) is called the Generalised Extreme Value distribution (GEV). In Equation (4), μ is the location parameter, σ is the scale parameter and ξ is the shape parameter. Estimating the parameters $\theta = \{\mu, \sigma, \xi\}$ of the GEV is done by means of maximum likelihood. The maximum likelihood equation is

$$L(X; \theta) = -n\, ln(\sigma) - (1 - \frac{1}{\xi}) \sum_{i=1}^{n} ln(y_i) - \sum_{i=1}^{n} y_i^{-1/\xi}, \tag{6}$$

where $y_i = 1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)$, $i = 1, \ldots n$ (Hosking, 1985).

4.2 Heavy tailed variables

This thesis deals with variables that have observations in the heavy tails and only take on positive values. Formally a random vector $\mathbf{x} \in \mathbb{R}_+^p = [0, \infty)^p$ is regularly varying if there exists a sequence $b_n \to \infty$ and a limit measure $v_{\mathbf{x}}$ for sets in $(0, \infty)^p$ such that

$$n\, pr(b_n^{-1}\mathbf{x} \in C) \xrightarrow{v} v_{\mathbf{x}}(C) \text{ as } n \to \infty, \tag{7}$$

where $\xrightarrow{v}$ denotes vague convergence (Resnick, 2007). It holds that $b_n = L(n)\, n^{1/\alpha}$ for some slow varying function $L(n)$ and $\alpha > 0$ is called the tail index of $\mathbf{x}$. Because of its definition, $v_{\mathbf{x}}(C)$ has the scaling property $v_{\mathbf{x}}(aC) = a^{-\alpha} v_{\mathbf{x}}(C)$ for an arbitrary set $C \subset (0, \infty)^p$ with $a > 0$. We fix $\alpha = 2$ as suggested by Cooley & Thibaud (2019). Due to the scaling property of $v_{\mathbf{x}}$, it is much more convenient to look at sets defined by polar coordinates. First define the unit ball

$$\mathbb{S}_+^{p-1} = \{\mathbf{x} \in \mathbb{R}_+^p : ||\mathbf{x}|| = 1\} \tag{8}$$

and

$$C(r, B) = \{\mathbf{x} \in \mathbb{R}_+^p : ||\mathbf{x}|| > r, \frac{\mathbf{x}}{||\mathbf{x}||} \in B\} \tag{9}$$

for $r > 0$ and a Borel set $B \subset \mathbb{S}_+^{p-1}$. From these definitions we can obtain that $v_{\mathbf{x}}(C(r, B)) = r^{-\alpha} H_{\mathbf{x}}(B)$, where $H_{\mathbf{x}}(B)$ is called the angular measure on $\mathbb{S}_+^{p-1}$. Note that the superscript of the unit ball is $p-1$ and not p, seeing that we only need the values of $p-1$ variables to know the values for all p variables as the norm must be equal to one and all values are non-negative.
So far we assumed that the available data satisfies the previously mentioned restrictions. For most data sets however this is not the case, so we need a general way of transforming any data set so that we can always inspect the extreme values of data sets.

4.2.1 Transformation of variables

We would like to model observations which are classified as extreme values. To model these observations we first need an expression for the distribution function of the extreme values. For this we assume a sequence of constants exists $a_j^n > 0$, $b_j^n \in \mathbb{R}$, $j = 1, \ldots, p$. We denote the data as $\mathbf{X} = (X_1, \ldots, X_p)$, where p is the number of variables and we have n observations. Now the convergence that must hold is

$$\lim_{n\to\infty} P\left(\frac{max_{i=1,\ldots,n} X_1^i - b_1^n}{a_1^n} \leq x_1, \ldots, \frac{max_{i=1,\ldots,n} X_p^i - b_p^n}{a_p^n} \leq x_p\right) = G(x_1, \ldots, x_p). \tag{10}$$

One transformation that we can use so that restriction 10 is satisfied is by transforming the data based on the probability of each value occurring. We transform the variables in both data sets using

the formula

$$\mathbf{Y} = \left(\frac{1}{\sqrt{-ln[F(X_1)]}}, \ldots, \frac{1}{\sqrt{-ln[F(X_p)]}} \right). \tag{11}$$

In Equation (11) the expression $F(X_j)$ denotes the marginal distribution of variable j. For this thesis we use the empirical distribution function as suggested by Janßen & Wan (2020). The empirical distribution function that is often used is $\hat{F}_{j,n}(t) = \frac{1}{n}\sum_{i=1}^{n} I[x_{ij} \leq t]$, $j = 1, \ldots, p$, where $I[x_{ij} \leq t]$ is the indicator function that has value one if the restriction between the brackets holds and zero otherwise and $t > 0$ is some threshold. However an issue arises when using this formula. By plugging in the maximum value into Equation (11), i.e. we have the value for the empirical cdf of $\hat{F}_{j,n}(max\{x_{1j}, \ldots, x_{nj}\}) = \frac{n}{n} = 1$, we get a transformed value of $\frac{1}{\sqrt{-ln(1)}} = \frac{1}{0}$, which is not defined. Therefore we use a slightly altered version of the empirical cdf function, where the denominator is $n+1$: $\hat{F}_{j,n}(t) = \frac{1}{n+1}\sum_{i=1}^{n} I[x_{ij} \leq t]$, $j = 1, \ldots, p$. Plugging the transformed sample $\mathbf{Y}$ in restriction 10 we get

$$\lim_{n\to\infty} P\left(\frac{max_{i=1,\ldots,n} Y_1^i}{n} \leq x_1, \ldots, \frac{max_{i=1,\ldots,n} Y_p^i}{n} \leq x_p \right) = G_0(x_1, \ldots, x_p). \tag{12}$$

For this new sample $\mathbf{Y}$ the convergence in restriction 10 holds and we can perform Extreme Value Analysis. The subsequent subsections talk about the methods that we perform. For all methods we assume that the data is transformed as described by Equation (11).

4.3 Spherical k-means

One assumption that we often make when analysing data is that all observations are independent and have similar characteristics. In more mathematical terms, we assume all observations to come from the same distribution with the same parameters, which we call iid. This assumption however may not always be reasonable. Data could be observed over a very long time period or other factors that are not observed in the data can have an (unobserved) effect on the data. In fact these differences might even be of significant interest as they they can help to explain the structure in the data. One straightforward and frequently used method to examine the differences between groups of data is called spherical k-means. This method is similar to the perhaps more well-known method called k-means. Spherical k-means starts off by defining a set of $k \in \mathbb{N}$ different groups called "clusters". As the name somewhat implies, our main interest is the centroid (mean) of each cluster. We denote the set of centroids as $A = \{a_1, \ldots, a_k\}$, $a_c \in \mathbb{R}^p$, $c = 1, \ldots, k$ (Buchta et al., 2012). The next step in performing spherical k-means is actually assigning observations to any of the clusters. Observations can be assigned to any cluster, but every observation is only assigned to one cluster (hard clustering). By simply looking at the data it is practically impossible to assign observations to clusters by hand, so we need an algorithm that optimally does this for us.

4.3.1 Assigning observations optimally

The way in which we optimise spherical k-means is that we set up clusters and assign observations to any cluster so that the distance between observations within the same cluster and their respective centroid is minimised, where the number of clusters k is specified beforehand (Likas et al., 2003). Combining what we so far know about spherical k-means with the fact that data is transformed as explained in Subsection 4.2.1, we thus need to solve the equation

$$W(A, P) = \int_{\mathbb{S}_+^{p-1}} \min_{\mathbf{a} \in A} d(\mathbf{x}, \mathbf{a}) P(d\mathbf{x}) \in [0, \infty), \tag{13}$$

where P is a probability measure and $d(\mathbf{x}, \mathbf{a})$ is a function that measures the distance between $\mathbf{x}$ and $\mathbf{a}$ (Janßen & Wan, 2020). Note that the value of $W(A, P)$ is strictly non-negative, as the data itself is also non-negative. So far we have talked about "distance" between points, but distance can be measured in various ways. The distance measure is the part where spherical k-means is different from "regular" k-means, as spherical k-means uses the cosine dissimilarity as distance measure instead of the Euclidean distance (Dhillon & Modha, 2001). The formula for the distance between observation $\mathbf{x}$ and the centroid of its cluster $\mathbf{a}$ using the cosine dissimilarity is

$$d(\mathbf{x}, \mathbf{a}) = d_\varphi(\mathbf{x}, \mathbf{a}) = 1 - cos(\mathbf{x}, \mathbf{a}) = 1 - \frac{\mathbf{x} \cdot \mathbf{a}}{||\mathbf{x}||_2 ||\mathbf{a}||_2}, \tag{14}$$

where $||.||_2$ denotes the Euclidean norm. Normalising all variables to have unit length so that the denominator in Equation (14) is 1, reduces the formula to $1 - \mathbf{x} \cdot \mathbf{a}$. For P in Equation (13) we substitute its sample version S_n, which assigns weight $1/n$ to every observation. We now obtain a set of sample means denoted A_k^n. Everything neatly summarised, assigning observations in our data optimally to clusters is done by solving

$$\begin{aligned} W(A_k^n, S_n) = & \min_{\mathbf{a}, \mu} \sum_{c=1}^{k} \sum_{i=1}^{n} \mu_{ic} \left(1 - \frac{x_i \cdot a_c}{||x_i||_2 ||a_c||_2}\right), \\ & s.t. \sum_{c=1}^{k} \mu_{ic} = 1,\ i = 1, \ldots, n, \\ & \mu_{ik} \in \{0, 1\},\ i = 1, \ldots, n,\ c = 1, \ldots, k, \\ & a_c \in \mathbb{R}^p,\ c = 1, \ldots, k, \end{aligned} \tag{15}$$

where $\mu_{ic} = 1$ if observation i is assigned to cluster c and 0 otherwise (Buchta et al., 2012).

4.3.2 Optimisation of k

Choosing the correct number of clusters in (spherical) k-means is very important for getting optimal results. In an ideal world we know the optimal numbers of clusters k in which to segment the observations beforehand. Unfortunately this is not the case in practice. We do know however that

increasing the number of clusters decreases the variance within clusters. By knowing this one might think to use as many clusters as possible, but increasing the number of clusters also makes the model more biased towards the used data set. Having a large number of clusters also ensures most clusters are sparse which makes it difficult to interpret the results. Thus, we need a sufficient number of clusters to ensure a small variance, but the number of clusters should not be too large so that the model is not biased towards the data and results are interpretable. The most straightforward way for finding the optimal number of clusters is to use some metric whose value we can compare for different values of k (Pham et al., 2005). The metric that one uses to find the optimal number of clusters is usually the distance measure that is used to optimally assign observations to clusters. We use the cosine dissimilarity measure, so our metric is Equation (14). We now construct a set containing different numbers of clusters $\{2, 3, 4, \dots\}$. Next for all number of clusters in the set we perform the spherical k-means algorithm as explained in the previous subsection. The next step is to set up a so-called elbow plot. This plot shows the number of clusters plotted against the value of Equation (14). As the name suggests, the line drawn between the points somewhat resembles an elbow and we look for a significant decline in the value of the metric. The point at which this happens signifies the optimal number of clusters. A nice example of an elbow plot for simulated data as well as the Matlab code is given by Bholowalia & Kumar (2014). Note that the number of clusters should always be at most equal to the number of observations, otherwise some clusters cannot contain observations and thus have no interpretation. In practice the numbers of clusters should even be smaller than the number of observations, because having as many cluster as observations means that every observation contains only one observation. This would not add anything to the interpretability of each cluster.

4.4 Principal Component Analysis

In unsupervised learning methods, the data we use often consists of many variables. Some variables in the data might be of greater importance than other variables. By simply looking at the data we cannot easily deduce which variables are most important. A frequently used method that looks at the relative importance of variables is called Principal Component Analysis (PCA). If we have a data set denoted $\mathbf{X} = (X_1, \dots, X_p)$, then we can construct the $p \times p$ covariance matrix Σ_X, where p refers to the number of variables present in the data. This matrix has the covariance between variable j and variable k on element σ_{Xjk}, where jk refers to the element on row j and column k, $j, k = 1, \dots, p$, $j \neq k$. The variance of variable j can be found on the diagonal element σ_{Xjj}, $j = 1, \dots, p$. Due to the specification of the covariance matrix it holds that this matrix is both symmetric and positive semi-definite (Yang & Berger, 1994). From this matrix Σ_X we can calculate all eigenvalues by solving the equation $det(\Sigma_X - \lambda I_p) = 0$ for the eigenvalues λ, where $det(U)$ refers to the determinant of matrix U and I_p denotes the identity matrix of size p. The corresponding eigenvectors are obtained by finding the null space of the matrix $\Sigma_X - \lambda I_p$ by means of row operations. For the mathematical details on eigenvalues and eigenvectors we refer to the

book by Johnson & Wichern (2014).
After having obtained each eigenvalue and eigenvector, we construct each eigenvalue-eigenvector pair, denoted $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_p, \mathbf{e}_p)$. Each eigenvector $\mathbf{e}_j$ is made up of the elements $\mathbf{e}_j = [e_{j1}, e_{j2}, \ldots, e_{jp}]'$, $j = 1, \ldots, p$, where ' refers to the transpose of a matrix or vector. As noted earlier, the covariance matrix Σ_X is positive semi-definite which means that for the eigenvalues the inequality $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ holds. Based on the eigenvalue-eigenvector pairs we can construct new principal components, where principal component j is given by the formula

$$PC_j = \mathbf{e}_j'\mathbf{X} = e_{j1}X_1 + e_{j2}X_2 + \cdots + e_{jp}X_p,\ j = 1, \ldots, p. \tag{16}$$

Every principal component is a linear combination of all original variables, so that the elements of an eigenvector show the importance of each variable for a principal component (Johnson & Wichern, 2014). This has the convenient added effect that components sometimes have a logical interpretation. Because the eigenvalues are ordered based on their values, we know that the formula

$$\hat{\lambda}_j = \frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k},\ j = 1, \ldots, p, \tag{17}$$

gives the relative importance of principal component j. We refer to $\hat{\lambda}_j$ as the relative eigenvalue. This formula is also ordered for all eigenvalues, which implies that a large value for Equation (17) means that the corresponding component is of great importance. For example a value of $\hat{\lambda}_j = 0.3$ for this formula means that the component to which this eigenvalue belongs explains 30% of the variability in the data.
Now we could construct all p principal components and try to explain our results by looking at the coefficients of every single principal component. This however may not be convenient nor even feasible if our data contains many variables. It is more useful to only examine the k most important principal components where we want k to be (much) smaller than the total number of variables p. Two ways are often practised to achieve this. First we could examine Equation (17) somewhat closer. As noted earlier, the value for $\hat{\lambda}_j$ can be interpreted as the percentage of the variance explained by component j. This also means that the sum of relative eigenvalues signifies the percentage of the variance that the corresponding components together explain. Based on this information one approach for determining the optimal number of components is to beforehand specify a percentage of the variance that we want to explain. Then we only look at the minimal number of principal components (ranked from largest to smallest) that we need to exceed this threshold (Valle et al., 1999). The threshold is usually 80 or 90%, dependent on the data used. We use a threshold of 80%.
The second method we can use to find the optimal number of principal components is by looking at what we call the scree plot. For this plot we look at the ordered eigenvalues and the number of components. The x-axis shows the component, while the y-axis shows the value for the eigenvalue λ. Now similar to the elbow plot for spherical k-means, we look for a sudden significant decrease in the value of the eigenvalue. This sudden decrease indicates that the optimal number of components

is reached (Kanyongo, 2005). To recap, the goal of using principal components is to find "new" variables that explain much of the variance in the data. These new variables are simply linear combinations of the original variables, which therefore maintain the structure of the data and hence everything discussed in this subsection also applies to Extreme Value Analysis. Also Cooley & Thibaud analyse the extremal dependence using eigendecomposition, but the matrices that one obtains simply contain the eigenvalues and eigenvectors. Henceforth the theory discussed in this subsection also holds in the context of extreme values.

4.4.1 Tail pairwise dependence

In the previous subsection we discussed how to obtain the eigenvalues and eigenvectors from a given covariance matrix. Subsequently we talked about how to interpret the results. The previous subsection however assumed that we already know how to estimate the covariance matrix, but in the context of Extreme Value Analysis this is not as clear. The main issue stems from the fact that we want to model pairwise dependence only for observations that are classified as extreme outliers. Using "regular" formulas to estimate the covariance matrix Σ_X break down in this context as it does not maintain the structure and nature of extreme values. From this point on we will refer to the matrix Σ_X as the tail pairwise dependence matrix, to generalise the concept of a covariance matrix. To understand the formula for the tail pairwise dependence matrix, we first look at the formula that is used in regular cases. The covariance between variable x and y is then defined by the formula

$$cov(x, y) = E[(x - E(x))(y - E(y))] = \int\int (x - E(x))(y - E(y))f(x, y)dxdy, \tag{18}$$

where $f(x, y)$ indicates the joint probability density function of x and y (Heij et al., 2004). Equation (18) is scale dependent and so it does not accurately capture the nature of the extreme values, as all these observations have large values for some variables. A naïve approach to this issue could be to simply normalise all variables. This would mean that all variables have the same scale and therefore Equation (18) is now applicable to this transformed data set. If we would analyse the entire data set then we could indeed do this, but for out goals this will not solve the issue. The main problem stems from the fact that we are conducting Extreme Value Analysis. In other words we are only interested in the behaviour of observations that lie within the extreme ends of the tail. Therefore we only inspect a small fraction of the data. Applying Equation (18) to these extreme values will not ensure that the nature of the extreme values is retained, and therefore we are not able to use Equation (18). Instead we need a formula so that the requirements discussed in Subsection 4.2 of regular variation hold. Larsson & Resnick (2012) defined an altered formula for the covariance matrix that uses an angular measure to incorporate the structure of extreme values in a bivariate context. Cooley & Thibaud (2019) extended on this by making the formula compatible for a continuous interval. The tail pairwise dependence between variable j and variable

k can be calculated by evaluating the formula

$$\sigma_{Vjk} = \int_{\Theta^{p-1}} \omega_j \omega_k dH_V(\omega),\ j,k = 1,\ldots,p. \tag{19}$$

In Equation (19), $\Theta^{p-1} = \{\omega \in \mathbb{R}^p : ||\omega||_2 = 1\}$ and H_V is what we call the angular measure of V, where V refers to the set of extreme values. The integral is taken over an interval of positive and negative numbers, while the data only contains positive numbers. We can solve this issue simply by redefining the interval over which we take the integral. The formula now becomes

$$\sigma_{Xjk} = \int_{\mathbb{S}_+^{p-1}} w_j w_k dH_X(w),\ j,k = 1,\ldots,p. \tag{20}$$

In Equation (20) $w_j = \frac{X_j}{||X_j||_2} \in \mathbb{S}_+^{p-1}$, $j = 1,\ldots,p$. Finally Equation (20) assumes that we know the true angular measure. Larsson & Resnick have shown that by substituting the angular measure H_X for an empirical estimate, Equation (20) simplifies to

$$\hat{\sigma}_{Xjk} = \hat{m} \int_{\mathbb{S}_+^{p-1}} w_j w_k d\hat{N}_X(w) \approx \frac{\hat{m}}{n_{exc}} \sum_{i=1}^{n} w_{ij} w_{ik} I[||x_i|| > r],\ j,k = 1,\ldots,p. \tag{21}$$

In Equation (21), $\hat{m}$ is an estimate for $H_X(\mathbb{S}_+^{p-1})$ and r is some threshold based on the data above which we define observations to be extreme outliers, for example the 95th percentile. The number of observations above the threshold r is mathematically given by the formula $n_{exc} = \sum_{i=1}^{n} I[||x_i|| > r]$, where $I[u]$ denotes the indicator function which has value one if statement u is true and zero otherwise. By first transforming the data so that every variable in the data has a common unit scale, the estimate for the angular measure $\hat{m}$ is simply equal to the number of variables, i.e. $\hat{m} = p$. We transform variables as described in Subsection 4.2.1.

4.5 Spherical k-principal-components

In the previous subsections we discussed two methods that can be applied to analyse extreme values. The first method is called spherical k-means, which optimally allocates observations to clusters. The second method is called Principal Component Analysis. The goal of PCA is to explain the variance in the data using a limited number of components. In spherical k-means we are interested in the means of all clusters, while for PCA we look at the values for each entry of a limited number of components. One recently introduced method that combines both methods is called spherical k-principal-components. Fomichov & Ivanovs (2020) prove the validity of this method by first examining what it means for clusters to be optimal. Recall that for the partition

$X = (x_1, \ldots, x_k) \subset \mathbb{S}_+^{p-1}$ where X satisfies restriction 10 it holds that

$$E\left(\underset{c=1,\ldots,k}{max}\{x_c'X\}\right) \leq E\left(\underset{c=1,\ldots,k}{max}\{a_c'X\}\right) \text{ and } E\left(\underset{c=1,\ldots,k}{max}\{(x_c'X)^2\}\right) \leq E\left(\underset{c=1,\ldots,k}{max}\{(\mathbf{e}_c'X)^2\}\right), \tag{22}$$

where a_c denotes the cluster mean of cluster c and $\mathbf{e}_c$ denotes the eigenvector corresponding to eigenvalue λ_c. To see how we can use these two results we inspect what both inequalities actually mean.

The first inequality refers to spherical k-means. Thus if we define the set of partitions $\{A_1, \ldots, A_k\} = \mathscr{P}_k \subset \mathbb{S}_+^{p-1}$, we can also write the left hand side of the inequality as

$$\underset{x_1,\ldots,x_k \in \mathbb{S}_+^{p-1}}{max} E\left(\underset{c=1,\ldots,k}{max}\{x_c'X\}\right) = \underset{\mathscr{P}_k}{max} \sum_{c=1}^{k} ||E(I[X \in A_i])||_2. \tag{23}$$

For the second inequality we can similarly write it out in terms of Borel sets instead of partitions, which gives us

$$\underset{x_1,\ldots,x_k \in \mathbb{S}_+^{p-1}}{max} E\left(\underset{c=1,\ldots,k}{max}\{(x_c'X)^2\}\right) = \underset{\mathscr{P}_k}{max} \sum_{c=1}^{k} \lambda_1(\Sigma_c). \tag{24}$$

The term Σ_c in Equation (24) refers to the tail pairwise dependence matrix of cluster c. Especially Equation (24) shows us a nice result. The equality tells us that the main interest lies in the largest eigenvalue of every covariance matrix. In other words we would like to set up the covariance matrix for every cluster, from which we only look at the largest eigenvalue and corresponding eigenvector. To do this we first evaluate the product between the centroids and the data, as shown in Equation (23) and for each row we store which column contains the highest value. Next we approximate the covariance matrix and find the largest eigenvalue and corresponding eigenvalue for each cluster, as shown in Equation (24). The most important eigenvectors of all clusters then give a new set of cluster means. We iterate over this procedure until a certain condition is met. In short the spherical k-principal-components algorithm goes as follows:

Algorithm 1: Spherical k-principal-components

Result: Optimal cluster centroids

Input: Normalised extreme values $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_{n_{exc}}) \subset \mathbb{S}_+^{p-1}$ and starting values $a_1, \ldots, a_k \in \mathbb{S}_+^{p-1}$;

Old cluster means $(\hat{a}_1, \ldots, \hat{a}_k) = (\infty_p, \ldots, \infty_p)$;

$diff = \sum_{c=1}^{k} ||\hat{a}_c - a_c||_2$;

$g = \mathbf{0}_{n_{exc}}$;

$\epsilon = 0.001$;

while $diff < \epsilon$ **do**

  $M = \hat{X}(a_1, \ldots, a_k)$ of size $(n_{exc} \times k)$;

  for $i = 1$ *to* $i = n_{exc}$ **do**

    $g_i = \underset{j=1,\ldots,p}{argmax}\{\hat{x}_{ij}\}$;

  end

  for $c = 1$ *to* $c = k$ **do**

    $\Sigma_c = \frac{1}{n_{exc}} \sum_{i=1}^{n_{exc}} (\hat{x}_c \hat{x}_c' I[g_i = c])$;

    Set a_c to first eigenvector of Σ_c;

  end

  $diff = \sum_{c=1}^{k} ||\hat{a}_c - a_c||_2$;

  Update old means $(\hat{a}_1, \ldots, \hat{a}_k) = (a_1, \ldots, a_k)$;

end

return Optimal cluster means $(a_1, \ldots, a_k)$

5 Results

This section discusses the main findings obtained from performing the proposed methods.

5.1 Finance data

The following subsections present the results obtained for the finance data set.

5.1.1 Spherical k-means

In this subsection we look at the results that have been obtained from performing the spherical k-means algorithm on the finance data set containing $n = 16,694$ observations. The goal of this method is to assign observations to clusters such that observations in the same cluster have similar characteristics. We then look at each cluster to see any possible patterns in the data. We use the 99th percentile to obtain $n_{exc} = 167$ extreme outliers. For these 167 observations we would like to determine how we can optimally assign them to clusters. We therefore first need to know into how many clusters we should segment the data. To answer this question we look at the summed

distances plotted against the number of clusters, also called the elbow plot. This plot is shown in Figure 1.

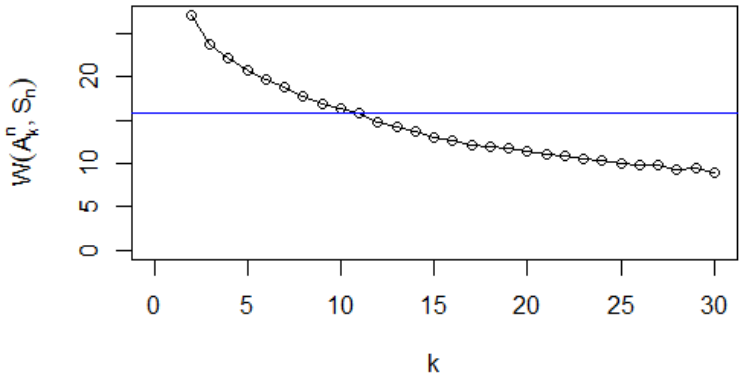


Figure 1: Number of clusters plotted against $W(A_k^n, S_n)$ for finance data.

The elbow plot for the finance data mostly shows a gradual decline for the values of $W(A_k^n, S_n)$ over all clusters. However the elbow plot seems to become most flat after the point of $k = 11$ clusters. Therefore we view $k = 11$ as being the optimal number of clusters and henceforth we will further examine the results associated with $k = 11$ clusters with the summed distances of $W(A_k^n, S_n) = 15.28621$.

To ease the interpretation of numbers we divide each centroid by the centroid of the variable which has the largest mean. Mathematically we thus perform the transformation

$$\left(\frac{a_1^c}{max_{j=1,\dots,p}\{a_j^c\}}, \dots, \frac{a_p^c}{max_{j=1,\dots,p}\{a_j^c\}} \right), c = 1, \dots, k. \tag{25}$$

The largest mean now has value 1 and all other values lie within the interval $[0, 1]$. The finance data set contains $p = 30$ variables, so it is difficult and highly inefficient to inspect every single cluster mean. If we would do that we need to inspect a total of $30 \times 11 = 330$ values. Instead it is more useful to look at the heatmap, which shows high cluster means thanks to light colours. Low cluster means have dark colours. By inspecting which variables have dark colours in the same cluster we know which variables arise in high quantities at the same time. The heatmap for all 30 variables of the finance data is shown in Figure 2.

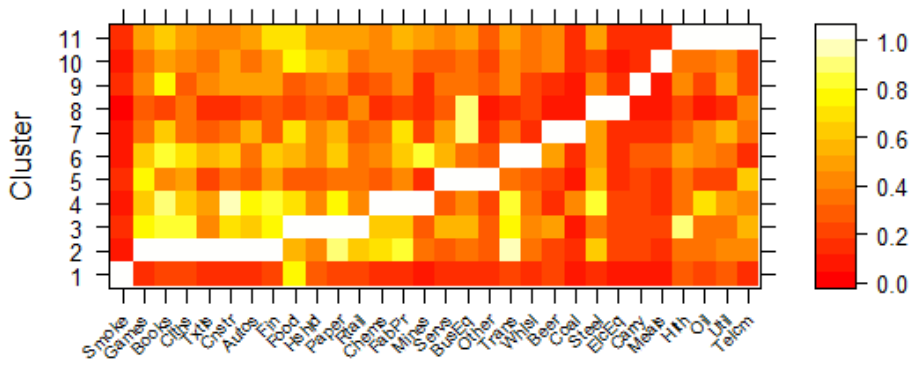


Figure 2: Heatmap of cluster centroids for finance data.

The most efficient way to examine the results is to look at each cluster separately and see which variables have a light colour, indicating that for this type of stock extreme losses have been frequently documented within this cluster.

- The first cluster is already remarkable, in the sense that only one stock has seen extreme losses; Smoke. The only other variable that has a somewhat light colour is Food. We thus conclude that cluster one shows extreme losses in businesses selling tobacco products and to a smaller extent food businesses. Moreover Smoke has a very dark colour in all other clusters so extreme losses in smoking mostly occur independently of other extreme losses. Especially the first observations seems logical, as tobacco products and food are both products that individuals consume on a daily basis.

- The second cluster shows joint high losses of the variables Games (anything recreational such as fishing), Books, Clths (clothes), Txtls (textiles), Cnstr (construction materials), Autos and Fin (financial institutions like banks). It is not surprising to find high losses for the variables Cnstr, Autos and to a lesser extent FabPr (fabricated products and machinery) in the same cluster as one needs any of the products to build the other. For example construction materials are needed to make machines, which we in turn need to make those construction materials as well as large products like cars. Furthermore we see extreme losses in textiles, which is also a material used to make other products. Therefore it is not a surprise to see extreme losses in Clths in the same cluster. The remaining variables Games and Fin seem somewhat out of place. These types of businesses do not seem to have connections with each other nor with the aforementioned businesses, but they may provide useful insides by performing further

research. In short cluster two contains losses in companies that provide some sort of material, as well as companies selling the products that need these materials.

- Cluster three shows joint extreme losses for the variables Food, Hshld (consumer products such as furniture or jewellery), paper (this includes object made of paper and wood, such as wooden furniture or pencils) and Rtall (retail) as well as Hlth (healthcare products) and Telcm (communication services such as phone or tv) to a lesser extent. This cluster mostly shows extreme losses for business dealing with products that people use on a daily basis. For example everyone has to eat food daily and most people use their phone or television every single day. This in turn explains why retail stores are in the same cluster, because people need to go to these types of stores in order to buy products that they use daily.

- Cluster four shows extreme losses for Cnstr, Chems (chemical products including paint and agricultural products), FabPr, Mines (mines for precious minerals). This cluster also contains less extreme but still significant losses in Oil and Util (utilities such as gas and electricity). Cluster four clearly contains extreme losses of companies which deal with anything that serves as some type of energy. This in turn makes it remarkable that extreme losses in Coal stocks are not present in the same cluster. Apart from energy businesses this cluster also contains businesses dealing with chemical products or construction products. These seem somewhat out of place in a cluster with energy companies, but conducting further research may lead to useful discoveries.

- Cluster five contains extreme losses for companies in the sectors Servs (service sector), BusEq (business equipment) and Other (sanitary services, air conditioning and irrigation systems). This cluster mainly shows extreme losses in the service sector, as all products in this cluster are used by companies in the service sector. Another loss worth mentioning is the losses in Telcm stocks, as businesses in the service sector need products that facilitate communication like phones to maintain contact with clients. Moreover service sectors, just like any other company, need business equipment. Finally the companies in the Other category can somewhat be viewed as being businesses in service sector themselves.

- Cluster six is only dominated by observations showing extreme losses in Whlsl (wholesale) or Trans (transport services) stocks. Combining this with the fact that Whlsl stocks do not show extreme losses in any other cluster, we conclude that extreme losses in the stocks of wholesale companies are always observed independently of other stocks, apart from Trans stocks.

- Cluster seven shows white for the variables Beer and Coal. Beer does not seem to have anything in common with Coal, so this cluster does not have a logical interpretation. The only conclusion that could be drawn from this cluster, is by combining cluster four and seven. As cluster four contains other energy related companies, we can see clusters four and seven combined as extreme losses of any energy providing company.

- Cluster eight only contains extreme losses for Steel (any product made of steel, e.g blast furnaces) and ElcEq (electrical equipment). It is reasonable to assume that electrical equipment and steel are used together when making various products, so it is no surprise to see these two variables being present in the same cluster. Both variables do not shows extreme losses in other clusters, so extreme losses in any of the two stocks always goes together with extreme losses in the other stock, independently of all other stocks.

- Cluster nine only shows extreme losses for Carry (aircraft, ship and railroad equipment). No cluster outside of cluster nine contains extreme losses for Carry stocks. It also should be noted that some extreme losses in Trans are observed in this cluster, which is logical as these types of companies go hand in hand with the companies that make transporting goods possible in the first place. All in all cluster nine shows extreme losses observed in the transport sector, mainly the construction of the vehicles and infrastructure used for transport.

- Cluster ten only contains observations that show extreme losses in Meals (restaurants and hotels). Similar to other clusters only containing one extreme loss, extreme losses for Meals stocks are not observed in any other cluster so extreme losses in stocks for this type of company are always observed independently. The only other variable in this cluster that is worth mentioning is Food, which is not surprising.

- Finally we examine cluster eleven. This cluster is dominated by observations showing extreme losses in Hlth, Oil, Util and Telcm. This cluster shows a similar result to cluster four, in the sense that extreme losses in some heavy industry stocks are jointly observed. Having extreme losses in Hlth and Telcm in the same cluster as extreme losses in Oil and Util is not an intuitive result, so further research is required to investigate whether there is an underlying connection between these types of stocks.

5.1.2 Principal Component Analysis

We examine the main results obtained from the Principal Component Analysis using the finance data set. Similar to the spherical k-means algorithm, we use the 99th percentile so that we are left with $n_{exc} = 167$ extreme observations. To see how many components we need to analyse, we first look at the number of components plotted against the value for the eigenvalues. Note that on top of the transformation from Equation (11), we have transformed the data to have Euclidean norm 1 for every variable, so all eigenvectors sum up to the number of variables. In other words $\sum_{j=1}^{p} \lambda_j = p$ with $p = 30$. The scree plot is shown in Figure 3.

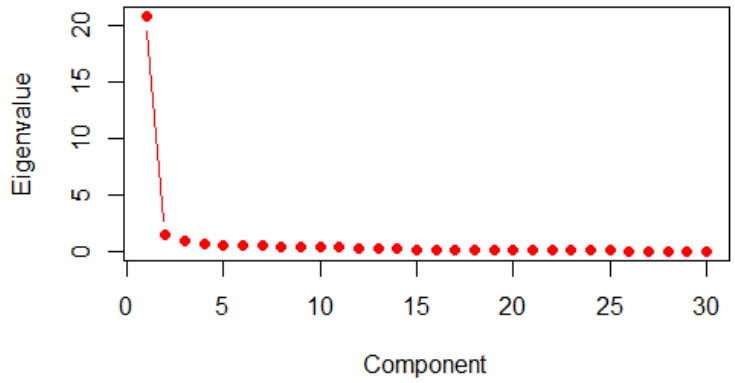


Figure 3: Scree plot finance data.

Looking at the scree plot we see a very significant drop in the eigenvalues after the first component. This means that the first component is by far the most important component in terms of explaining the variance in the data. The total number of variables is 30, so only including the first principal component would not be sufficient. To see how many components we should analyse we zoom in onto the eigenvalues for components 2 to 30. This is shown in Figure 4.

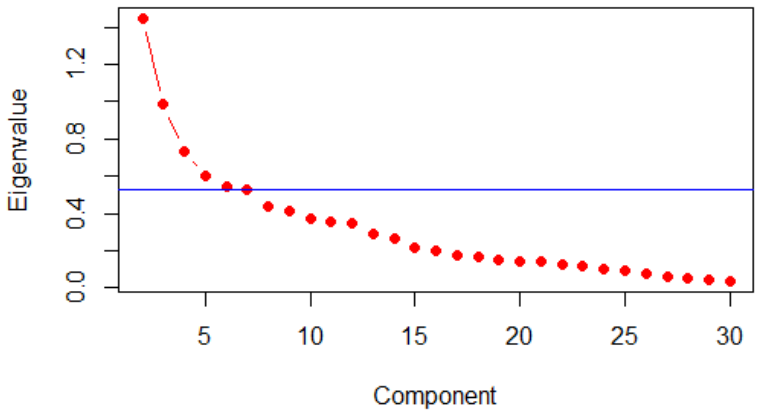


Figure 4: Scree plot finance data $\lambda_2, \ldots, \lambda_{30}$.

The scree plot for the finance data excluding the first component gives us a better idea of how many components to include. For the first few components we see a steep drop in the eigenvalue, after which the value does not change much starting at component seven. The rest of Figure 4 does

not show a point that seems optimal, so the scree plot tells us that analysing seven components is optimal. We also look at the cumulative variances to see whether seven components is indeed the optimal number for the finance data set. The cumulative variance after including any number of principal components as well as the prespecified threshold of 80% are shown in Figure 5.

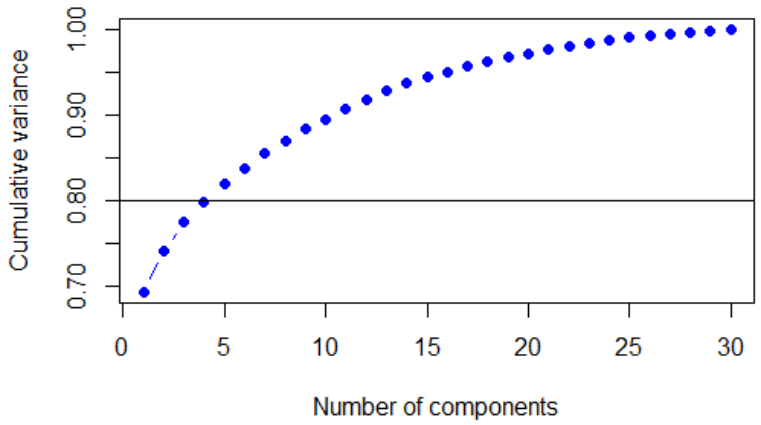


Figure 5: Cumulative variance plotted against number of components finance data.

First of all the figure for the cumulative variance shows something that we already noticed in the scree plot from Figure 3, namely that the first component already explains much of the variation present in the finance data. Secondly we see that the threshold of 80% is surpassed after including five principal components, which is two components less than the number of components which was deemed optimal by the zoomed in scree plot from Figure 4. Note that the goal of performing PCA is to obtain a certain number of principal components that explain most of the variation in the data set, where we want the number of components to be as few as possible. Therefore it would seem logical to include five instead of seven components, especially seeing that five components already explain 80% of the variation in the finance data. Therefore we will further inspect the result for the first 5 principal components. The finance data set contains a total of $p = 30$ variables, which gives us $5 \times 30 = 150$ values to look at. It would therefore be inefficient to inspect every single value and compare them. For this reason we visualise the values similar to what we did with the heatmap. We first transform the values of the eigenvectors using Equation (2), so that the numbers are easier to interpret. Values that were previously negative now lie below the threshold of $ln(2) \approx 0.69$, while previously positive values lie above this threshold. The values for the first five eigenvectors are visualised in Figure 6.

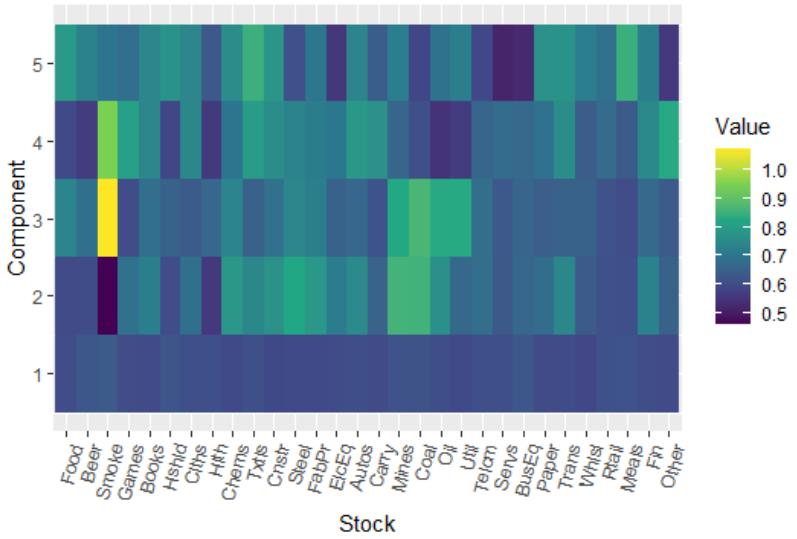


Figure 6: Colour scale first five eigenvectors finance data.

The colour scale represents extreme negative losses using a light yellow colour, while stocks that show no sign of an extreme loss are coloured dark and blue. Similar to each cluster from the spherical k-means algorithm, we inspect each component separately.

- The first component does not show any important results, as all values fall within a small interval. We can therefore interpret component one as being the general loss of stocks.

- The second component shows a clear contrast between two significant industries. First this component has the highest value for the variables Steel, Mines and Coal, which all correspond to companies that are part of the heavy industry. The smallest values are observed at the variables Smoke, Food, Beer, Consumer Goods and Healthcare. These variables resemble products that individuals comsume. Therefore this eigenvector shows the contrast between heavy and non-heavy industries.

- The variable that by far has the highest value in component three is Smoke. We also see significantly high values in the variables Mines, Coal, Oil and Util, which are all heavy industries as well as companies that supply some sort of energy/power. The lowest values are observed at the variables Servs and BusEq, which both correspond to companies in the service sector. This component thus shows the extreme losses in smoking companies as well as heavy industries, mitigated by losses in stocks service providing businesses.

- The penultimate component number four again has the largest value for Smoke, albeit smaller than the value in component three. We also see high values for the transport variables Auto, Carry and Trans as well as for the recreational variables Games, Books, Clothes and Paper. The extreme losses in any of the aforementioned businesses are contrasted by the heavy

industries, as the variables Mines, Coal, Oil and Util all have small values. Another sector that has small values for its variables is consumables, as the variables Food, Beer, Healthcare and consumer goods have dark blue colours. It is notable that the value for Smoke is much larger than that of the other consumables. Component four thus contrast smoking, recreational and transport companies with heavy industries and consumer goods (excluding smoking).

- Finally component number five has the highest values for Food, Beer, Smoke, Consumer Good and Meals. Another strangely placed variables with a high value is Textiles. Textile companies as well as companies in the consumable/food industries are contrasted by companies that are specialised in either Healthcare, Steel, Electrical Equipment, Telcom, Service or Business Equipment. This component seems most out of place of all components. For example many consumer goods have high values, while Healtcare has a small value. Also apart from Steel and Electrical Equipment, the variables with small values do not seem to have anything in common. Further research is required to determine whether any underlying connections are present.

5.1.3 Spherical k-principal-components

In this part we inspect the final results obtained for the finance data set by performing the Spherical k-principal-components algorithm. As noted in Subsection 4.5 we need a starting value for the cluster means as well as a fixed value for the number of clusters k. To not complicate things too much we use the same k as we deemed optimal from spherical k-means, which means we use $k = 11$. Moreover using the same number of clusters increases the likelihood of obtaining similar results. We also use the 99th percentile to obtain $n_{exc} = 167$ extreme values. We obtain results using two different starting values. First we draw random values from a uniform distribution and normalise these values so that all clusters have a mean with norm 1. Secondly the output of spherical k-principal-components can be seen as cluster means, similar to the output of spherical k-means. Therefore we also use the cluster means that were found by spherical k-means as starting values. Because we obtain cluster means, we will once again inspect the results by means of a heatmap, where all cluster means are normalised using Equation (25). The heatmap using starting values drawn from a uniform distribution using the seed 12345 is shown in Figure 7.

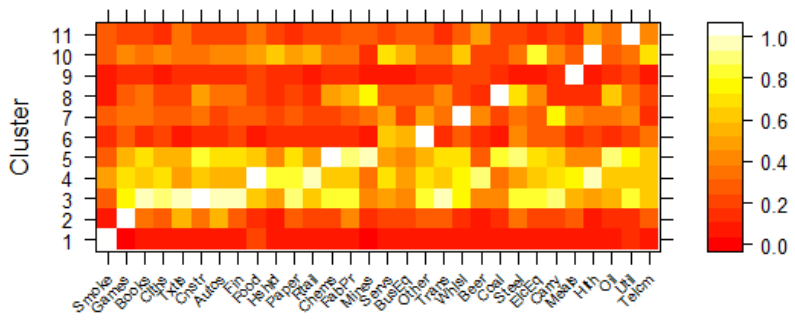


Figure 7: Heatmap of cluster centroids for finance data using random uniform starting values.

The results for spherical k-means were already discussed in detail in Subsection 5.1.1 and the results are partially overlapping, so discussing every single cluster is redundant. Therefore we will only touch on significantly different results. One thing that immediately stands out when looking at the heatmap from spherical k-principal-components is that many clusters only show a light colour for one variable. Clusters one and nine especially show this, as these clusters show white for only one variable while all other variables are practically inconsiderable. Extreme losses in Smoke, Games, Healthcare, Util, Wholesale, Meals and Other all seem to be observed independently of all other losses if we exclude clusters three, four, five and eight. Clusters three, four, five and eight however show more concurrent results with those obtained by spherical k-means. First of all cluster four shows joint extreme losses in Food, Beer, Healthcare and Retail. The first three of those variables are products that individuals consume, while seeing Retail in the same cluster is not surprising as one needs those store to actually buy these products. These losses are mitigated by Steel, Mines and Coal, which are all part of heavy industries. Cluster five shows extreme losses in heavy industries, namely Steel, Mines, Coal, Oil and Util. Other heavy losses in this cluster are Chems, FabPr, Electrical Equipment and Autos which could be linked to the previously mentioned industries but further research is required to look into that connection. Losses in this cluster are mainly mitigated by extreme losses in Beer and Smoke, which are consumed products.

We now look at the heatmap obtained by using the spherical k-means clusters as starting values. The heatmap can be seen in Figure 8.

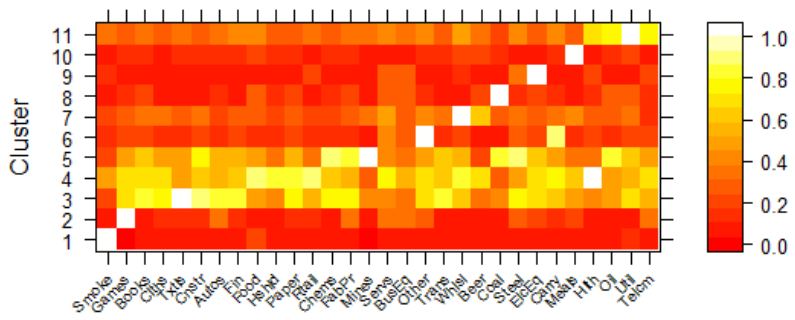


Figure 8: Heatmap of clusters centroids for finance data using spherical k-means starting values.

The heatmap using the starting values as obtained by the spherical k-means algorithm show, similarly to the previous heatmap, that most clusters are dominated by one variable. The number of clusters that are dominated by one variable however is lower, while the clusters that are dominated by one variable show an even more extreme focus on one variable. In this case extreme losses of Smoke, Games, Electrical Equipment, Coal and Meals dominate a separate cluster. Cluster eleven shows extreme losses in only two heavy industries, while remarkably being mitigated by extreme losses in Mines and Coal. Cluster four also shows unintuitive results as extreme losses in Steel, Oil and Utilities contrast losses in Mines and Coal. The same cluster also shows extreme losses in consumed products such as Food, Beer and Smoke. Cluster five perhaps produces the most logical results, as this cluster shows joined extreme losses in all heavy industries. Cluster three shows a similar result.

5.2 Winter pollution data

The subsequent subsections discuss the results obtained for the winter pollution data set.

5.2.1 Spherical k-means

We now look at the results obtained by performing spherical k-means on the winter pollution data set as described in Subsection 4.3. We first normalise all $n = 532$ observations. We use the 90th percentile and obtain $n_{exc} = 54$ extreme values. Then we would like to determine the optimal number of clusters by using the so-called elbow plot, which is shown in Figure 9.

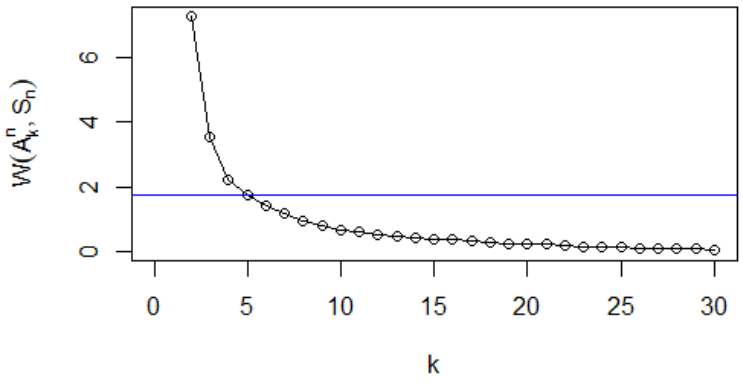


Figure 9: Number of clusters plotted against $W(A_k^n, S_n)$ for winter pollution data.

The elbow plot shows a steep decline in the differences for the first few clusters, after which the decline becomes insignificant from $k = 5$ clusters onward. Therefore we will further inspect the results obtained for $k = 5$ clusters with a value of $W(A_k^n, S_n) = 1.740996$. We normalise the cluster means using Equation (25) so that the results are more straightforward to interpret. The normalised cluster means for all variables are shown in Table 3.

Table 3: Normalised means for all variables in the winter data using $k = 5$ clusters.

Cluster	O_3	NO_2	NO	SO_2	PM_{10}
5	0.118	0.475	0.266	0.180	1
4	$0.697 \cdot 10^{-1}$	0.258	0.142	1	0.210
3	$0.731 \cdot 10^{-1}$	0.434	1	0.226	0.284
2	0.135	1	0.601	0.277	0.511
1	1	0.166	$0.982 \cdot 10^{-1}$	$0.722 \cdot 10^{-1}$	0.137

The normalised cluster means can be nicely visualised using a heatmap. This is shown in Figure 10.

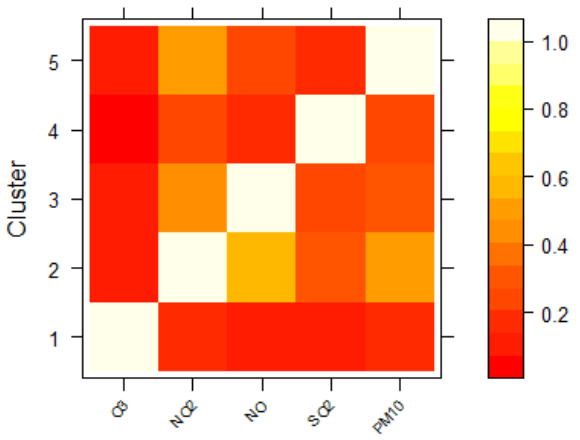


Figure 10: Heatmap of cluster centroids for winter pollution data.

The heatmap shows that each cluster mainly consists of observations that have a high pollution for just one of the five types. One notable exception to this observation is cluster five and to a lesser extent cluster one. For cluster five we see a high concentration of NO_2, while also observing a significant concentration of both NO and PM_{10}. Cluster one has a high concentration of PM_{10} and less but significant concentration of NO_2. Seeing this we can conclude that NO, NO_2 and PM_{10} can jointly occur in high concentrations, while high concentrations for any of the other types of pollution are observed individually during the winter months.

5.2.2 Principal Component Analysis

In this subsection we examine the main results obtained after carrying out Principal Component Analysis for the extreme values in the winter data set. We again use the 90th percentile to get $n_{exc} = 54$ extreme values. As described in the Methodology we first look at the scree plot and try to find a sudden drop in the eigenvalues. The scree plot for the winter data set is shown in Figure 11.

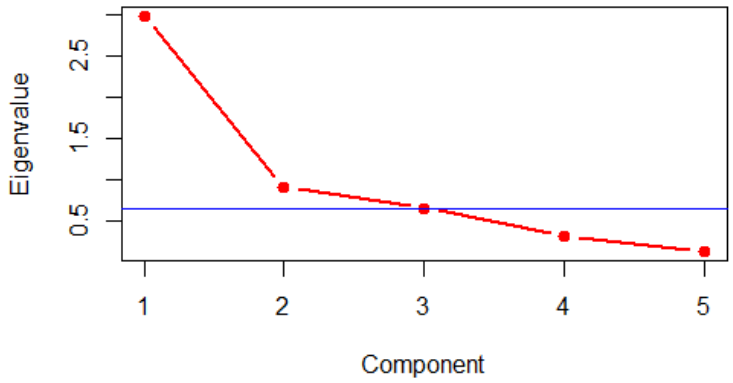


Figure 11: Scree plot winter pollution data.

In Figure 11 we see the number of components being plotted against the values for the eigenvalues. A sudden and significant drop is shown when we go from one to two and from two to three principal components, so three principal components in this context is considered optimal. Second we look at the values of the eigenvalues and the percentage of the variance that their corresponding components explain. Eigenvalues are shown in Table 4.

Table 4: Eigenvalues winter pollution data

Component	Eigenvalue	Variance	Cumulative variance
5	0.131	2.6%	100%
4	0.315	6.3%	97.3%
3	0.657	13.1%	91.1%
2	0.908	18.2%	77.9%
1	2.989	59.8%	59.8%

Inspecting Table 4 shows that we exceed the threshold of a cumulative variance of 80% when including three components, which is the same conclusion as the one drawn from the scree plot. Therefore we will further inspect the eigenvectors corresponding to the three most important principal components. Note that all eigenvectors sum up to the number of variables, i.e. $\sum_{j=1}^{p} \lambda_j = 5$, because we have pre-processed the data so that all variables have Euclidean norm 1. The eigenvalues from Table 4 may not exactly sum up to 5 due to the rounding of numbers. The same argument goes for the variances which should sum up to 100%. Finally all eigenvectors are divided by its absolute maximum value, which is shown in Equation (25) but inserting an absolute sign in every element of the denominator. This is mathematically correct as a linear combination of any eigenvector is also an eigenvector of the matrix. Also we are only interested in difference within one eigenvector, not

between eigenvectors. The values for all eigenvectors are shown in Table 5.

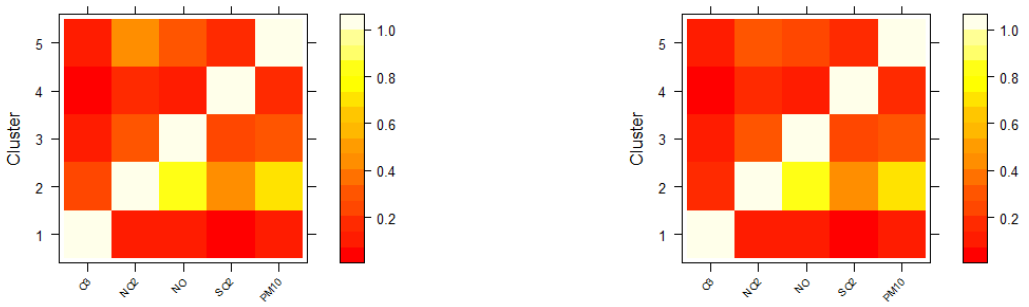
Table 5: Eigenvectors winter pollution data

Component	O_3	NO_2	NO	SO_2	PM_{10}
5	$-0.379 \cdot 10^{-2}$	1	-0.782	$-0.194 \cdot 10^{-1}$	-0.225
4	$-0.342 \cdot 10^{-1}$	-0.330	-0.710	$-0.287 \cdot 10^{-1}$	1
3	$0.432 \cdot 10^{-1}$	-0.259	-0.276	1	-0.277
2	1	-0.139	-0.146	-0.152	-0.116
1	0.502	0.970	0.931	0.763	1

For the eigenvector e_1 corresponding to the largest eigenvalue, there is not one value that is much further away from zero than all other values. The variable PM_{10} has the largest absolute value, but NO_2 and NO with values of 0.970 and 0.931 respectively are not far off. Using this information combined with the fact that all entries are positive the conclusion of this eigenvector is that e_1 shows the general pollution, mostly caused by PM_{10}, NO_2 and NO.
For the second eigenvector e_2 we see one absolute value that is significantly larger than all other values. For this eigenvector the value of 1 corresponding to the variable O_3 is the largest value. All other values are much smaller, so the second eigenvector signifies the pollution that only has extreme levels in O_3
Finally we look at the third eigenvector e_3. For this eigenvector again one value is much further away from zero than all other values. The variable SO_2 has the largest absolute value. The value that is second furthest away from zero (after 1) is -0.277, so we can safely assume that only SO_2 is relevant for this component. The conclusion of this eigenvector is that e_3 shows the extreme pollution caused by SO_2.

5.2.3 Spherical k-principal-components

This subsection touches on the results obtained by the spherical k-principal-components algorithm performed on the winter pollution data set. As explained in Subsection 5.1.3 we use two different starting values. The value for k is set to the same value as deemed optimal by the spherical k-means algorithm, so $k = 5$. Similar to the other two algorithms we use the 90th percentile on this data set so that we have $n_{exc} = 54$ extreme observations. The heatmap visualising the cluster centroids for the winter data using random uniform starting values is shown in Figure 12a, while the heatmap using the spherical k-means starting values is shown in Figure 12b.



(a) Heatmap of cluster centroids for winter pollution data with random uniform starting values.

(b) Heatmap of cluster centroids for winter pollution data with spherical k-means starting values.

Figure 12: Heatmaps of cluster centroids for winter pollution data after performing spherical k-principal-components.

The two heatmaps shows similar results. Almost every cluster is dominated by one type of pollution, with the main exception being cluster five in both cases. The conclusion from both heatmaps is thus that apart from NO_2, every type of pollution is observed independently of every other type. Extreme levels of NO_2 will also show less but still high levels of NO and PM_{10}. The two Figures 12a and 12b are almost identical. The only difference is the slightly darker shade for some variables within cluster four in Figure 12a compared to Figure 12b.

5.3 Summer pollution data

In the upcoming subsections we look at the results obtained by performing all proposed methods on the summer pollution data set.

5.3.1 Spherical k-means

We now look at the results obtained by performing the spherical k-means algorithm on the summer pollution data set as described in Subsection 4.3. We first normalise all $n = 578$ observations. Using the 90th percentile we are left with $n_{exc} = 58$ extreme values. Then we would like to determine the optimal number of clusters by using the elbow plot. Figure 13 shows the numbers of clusters being plotted against the value for Equation (15) for $k = \{2, 3, \ldots, 29, 30\}$ number of clusters.

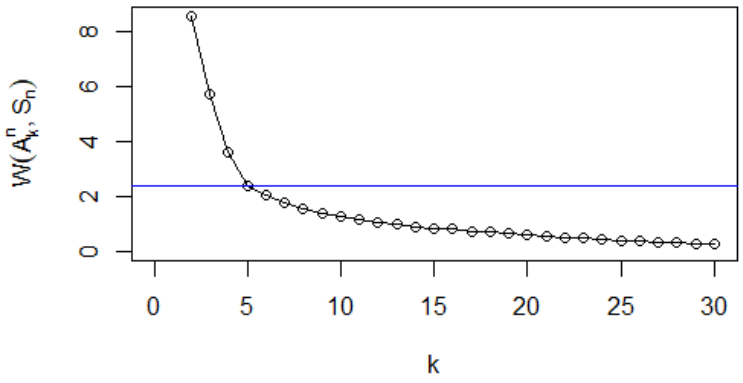


Figure 13: Number of clusters plotted against $W(A_k^n, S_n)$ for summer pollution data.

In Figure 13 we see a steep decline in the value for the optimisation function for the value $k = 2$ up to $k = 5$. After this the decline rapidly stagnates and the line drawn through the points almost becomes flat. Seeing this we view $k = 5$ with $W(A_k^n, S_n) = 2.089766$ as the optimal number of clusters, so we will further inspect this number of clusters.
First we look at the set of means denoted $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_5\}$. To ease the interpretation of numbers we divide each mean by the mean of the variable which has the largest mean. Mathematically we see this in Equation (25).

Table 6: Normalised means for all variables in the summer data using $k = 5$ clusters.

Cluster	O_3	NO_2	NO	SO_2	PM_{10}
5	0.190	0.223	0.196	0.250	1
4	0.213	0.397	0.230	1	0.355
3	0.213	0.495	1	0.269	0.338
2	0.377	1	0.397	0.243	0.355
1	1	0.296	0.125	0.259	0.303

The heat map for the summer pollution data that visualises the values in Table 6 is shown in Figure 14.

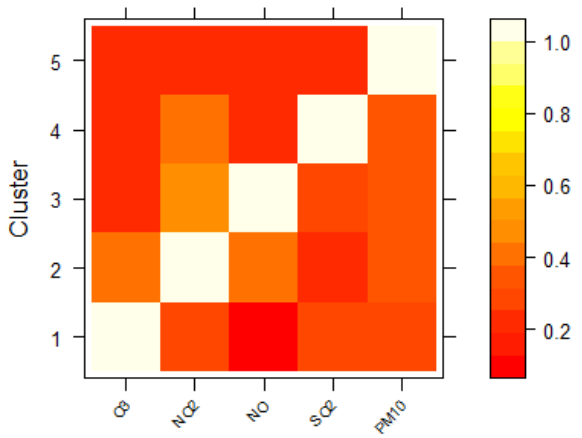


Figure 14: Heatmap of cluster centroids for summer pollution data.

We inspect each cluster in the heatmap separately. A white square indicates that the corresponding variable is largely represented in the cluster, while darker squares indicate that the variable is sparsely represented in the cluster. For each cluster (i.e. each row) we see a large concentration of one of the five types of pollution, while all other types have a low concentration. In other words every cluster mainly consists of observations that have high pollution of the same type and only one type of pollution. This hints at the fact that pollution of all types are observed independently of each other during the summer months.

5.3.2 Principal Component Analysis

The summer pollution data set has a total of five variables, so there are also five possible components to analyse. We use the 90th percentile to obtain $n_{exc} = 28$ extreme values. To determine the optimal number of components we both look at the first differences in the eigenvalues and the percentage of the variance they explain. We start off by looking at the first differences with the help of plotting the values for the eigenvalues against the number of components. This plot is shown in Figure 15.

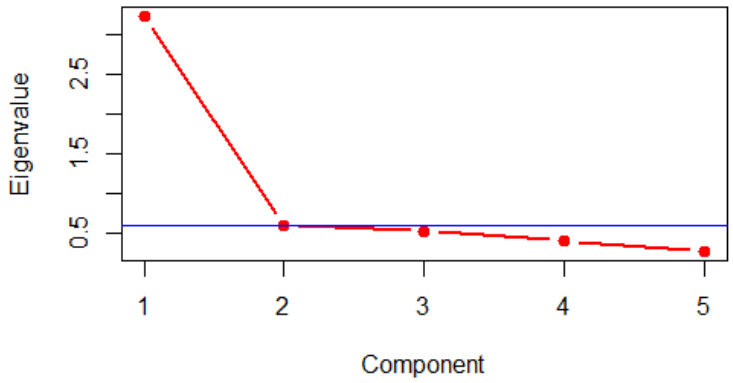


Figure 15: Scree plot summer pollution data.

A very steep decline in the eigenvalues is shown when going from one to two components. After this point the eigenvalues do not seem to go down very quickly anymore and the line stagnates. Therefore Figure 15 tells us that two components is optimal for this data set. To verify or contradict this statement we also look at the percentage of the variance explained by the eigenvalues. All eigenvalues and their corresponding percentages are shown in Table 7.

Table 7: Eigenvalues summer pollution data

Component	Eigenvalue	Variance	Cumulative variance
5	0.263	5.3%	100%
4	0.395	7.9%	94.7%
3	0.518	10.4%	86.8%
2	0.586	11.7%	76.5%
1	3.238	64.7%	64.7%

We first look at the individual percentages shown in Table 7. They tell us what we have already seen in Figure 15, namely that the first component explains much of the variance in the data, while the other components explain the rest of variance in similar quantities but much less than the first component. We now look at the cumulative variances, in which we see that the threshold of 80% is only exceeded after including a total of three principal components. This however contradicts to the observation that was made previously. Still the second and third principal component both explain more than 10% of the variance in the data, therefore it seems logical to also include them. Three components is thus viewed as optimal. This data set only contains $p = 5$ variables, so we do not obtain many values and we can inspect every number. All eigenvectors and their corresponding values can be found in Table 8.

Table 8: Eigenvectors summer pollution data

Component	O_3	NO_2	NO	SO_2	PM_{10}
5	-0.438	1	-0.751	$-0.877 \cdot 10^{-1}$	$0.702 \cdot 10^{-1}$
4	$-0.419 \cdot 10^{-1}$	-0.229	$-0.996 \cdot 10^{-1}$	-0.753	1
3	-0.734	-0.531	-0.343	1	0.567
2	1	-0.258	-0.948	0.235	$0.652 \cdot 10^{-1}$
1	0.805	1	0.856	0.853	0.991

Similar to each cluster in spherical k-means we divided each value by the absolute largest value in its respective component.

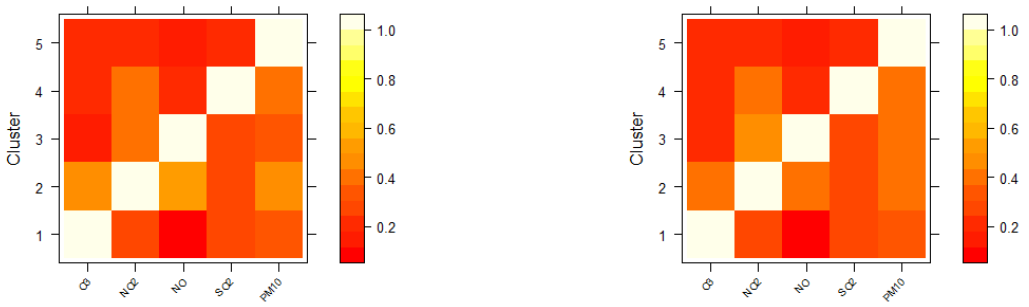
We start off by looking at the eigenvector corresponding to the largest eigenvalue. This eigenvector, denoted e_1, has no variable that immediately stands out. All values have the same sign (i.e. all are positive) and the difference between the largest and smallest value is less then 0.2. Therefore this variables shows general pollution.

The second eigenvector e_2 is, unlike e_1, mainly driven by two variables. This eigenvector is most heavily influenced by the variable O_3. Secondly, but to a lesser extent, the eigenvector is influenced by the variable NO seeing that this variable has a value of -0.948. The eigenvector e_2 thus corresponds to the pollution caused by PM_{10}, while being offset by the pollution caused by O_3 and vice versa.

Finally we look at the third eigenvector e_3. This eigenvector is distinct from the two previously discussed eigenvectors in the sense that only one value is far away from zero. The variable with the largest absolute value is SO_2, while the second most important is O_3 with a value of -0.734. The third eigenvector can therefore be seen as a variable that shows extreme pollution in SO_2 which is partially being offset by extreme levels of O_3.

5.3.3 Spherical k-principal-components

In the final part of the Results section we show and inspect the results obtained by performing spherical k-principal-components for the summer pollution data. We explained in Subsection 5.1.3 that we use two starting values and assume k to be the same as obtained by spherical k-means, so $k = 5$. We use the 90th percentile and are left with $n_{exc} = 58$ observations classified as extreme values. Figure 16a shows the heatmap using random starting values and Figure 16b shows the heatmap using spherical k-means starting values.



(a) Heatmap of clusters for summer pollution data with random uniform starting values.

(b) Heatmap of clusters for summer pollution data with spherical k-means starting values.

Figure 16: Heatmaps of cluster centroids for summer pollution data after performing spherical k-principal-components.

The heatmaps in Figure 16a and Figure 16b are very similar. In both heatmaps each cluster is dominated by one variable while other variables in the same cluster are insignificant. The conclusion of both heatmaps is that extreme levels in any type of pollution is always observed independently of other types of pollution. The only difference that one could mention between the two heatmaps is that the shades of some variables are slightly different, but they do not seem significant enough to change the conclusion.

6 Discussion and Future Research

In the previous section we discussed the results obtained by performing the proposed methods for all data sets. Now we would like determine what all results per data set have in common. At first glance (spherical) k-means and Principal Component Analysis seem like two completely different methods. Any form of k-means tries to optimally allocate observations to clusters, ensuring that similar observations end up in the same cluster. On the other hand PCA is a dimension reduction algorithm that tries to explain most of the variation in the data by constructing "new" components that are linear combinations of all variables in the data. However the two methods are very similar and one could even argue that both methods try to accomplish the same thing. A prominent paper that studies the relationship in depth is the paper written by Ding & He (2004). Ding & He prove that the analogy lies within the cluster means of the k-means algorithm. In short Ding & He state that the cluster means of each cluster have a similar interpretation to the values of a corresponding eigenvector. This means that if two variables have a high mean in a cluster, there is likely to be an eigenvector that has a high value for the same two variables. Spherical k-means is almost the same algorithm as "regular" k-means, apart from the fact that spherical k-means uses the cosine dissimilarity instead of the Euclidean distance. Therefore we can still use the result pioneered by Ding & He to compare the eigenvectors with the cluster means for each data set. We start the

analysis with the finance data set, as this was the first data set for which we performed the proposed methods. We thus compare the heatmap in Figure 2 with the colour scale in Figure 6. As mentioned earlier the third algorithm that we performed called spherical k-principal-components produces results that can be seen as cluster means. Therefore we can subsequently compare the spherical k-means and PCA results with the heatmaps in Figure 7 as well as Figure 8.

It should first be noted that the number of clusters that we deemed optimal for the spherical k-means is larger than the number of principal components we inspected. Therefore we cannot find a one-to-one correspondence of clusters and components. By inspecting each cluster and component, we indeed see some similarities. For example we see a cluster that is dominated by the variable Smoke, while we also see a component that is almost the same is simply the variable Smoke. Moreover we see two clusters that mainly contain extreme losses of stocks that belong to power supplying companies, while also observing two components that are a linear combination of several power supplying stocks. There however also seem to be some dissimilarities. For example the aforementioned cluster dominated by the variable Smoke also shows a relatively light colour for Food, while the analogous component has a positive but small value for Food and high values for heavy industries. The spherical k-means algorithm also produced a cluster dominated by Beer, while no principal component has an extremely high value for this variable. The differences might be a result of the fact that we do not have a similar number of clusters and components. Ding & He argue that k cluster are somehow linked to the first $k-1$ principal components, but further research is required to see whether this applies to this data and these Extreme Value Analysis methods. Now we look at the results obtained from the spherical k-principal-components algorithm, which perhaps show the most distinct results. First of all the heatmaps using two different starting values differ among each other. The most significant similarity is that most clusters are dominated by one variable, but the variables that dominate a cluster do differ somewhat. Also, the heatmaps when using spherical k-means results as starting values do not seem reliable. The heatmaps for example show that extreme losses in some heavy industries contrast extreme losses in other heavy industries. Therefore we will only compare the heatmaps using random uniform draws as starting values with the other results.

The heatmap in Figures 7 shows some similarities to both PCA and spherical k-means, mainly that extreme losses in heavy industries coincide with each other and one cluster is dominated by the variable Smoke. However there is one big difference between the spherical k-principal-components results and the spherical k-means and PCA results. The main difference is that in spherical k-principal-components results many clusters are dominated by one variable and not many results are intuitive. This could have two causes. Firstly using different starting values can produce very different clusters, as shown by the obtained results. Secondly we assumed k to be equal for spherical k-means and spherical k-principal-components. Future research could follow up on these two points by optimising spherical k-principal-components using different starting values as well as different values for k. Secondly we compare the results obtained for the pollution data sets. We compare

the centroids in Figure 10 with the numbers from Table 5 as well as the centroids from Figure 14 with the eigenvectors from Table 8. Again there is no one-to-one correspondence of cluster and components, but we do see some similarities. The centroids for the winter data show that extreme pollution for most types of pollution are independent of other types, while extreme levels of NO_2 will also see less but still significantly high levels of NO and PM_{10}. If we then look at the eigenvectors we see a similar figure. Each component is dominated by one variable, while component one has the highest value for PM_{10} and significantly high absolute values for NO_2 and NO. The heatmap for the summer pollution data shows a similar figure. In this case each cluster is fully dominated by one type of pollution, meaning that extreme levels of pollution for any type will occur independently of other types of pollution. Also cluster three shows that extreme levels of NO will also see high levels of NO_2. We see a similar figure in component one, which has the highest value for NO_2 and a high positive value for NO. However the second largest value of this component corresponds to the variable PM_{10}, while the value of NO is only the third largest. Finally we compare all these results to the heatmaps corresponding to spherical k-principal-components. The winter heatmaps are shown in Figures 12a and 12b, while the summer heatmaps can be found in Figures 16a and 16b. The heatmaps per data set are practically identical, while the heatmaps for different data sets differ somewhat. All heatmaps show the same results as obtained by spherical k-means. This means that during the winter extreme levels of pollution are observed independently and extreme levels of NO_2 will also see less but still high levels of NO and PM_{10}. In the summer months most types of pollution are observed independently of other types of pollution, while extreme levels of NO_2 will also see high levels of NO. From these findings we conclude that again there are many similarities between the numbers obtained by spherical k-means, PCA and spherical k-principal-components but also a notable number of differences. However for this data the results obtained by spherical k-means and spherical k-principal-components are almost the same. Still one must inspect all results to come to a definitive conclusion. Also the results for the winter and summer pollution data differ significantly, so it is indeed advised to expect both data sets separately.

The conclusion on the validity of all methods is that spherical k-means, PCA and spherical k-principal-components all produce logical and reliable results. However the results are not always identical. Spherical k-principal-components seems to work well when having a small number of variables. For the pollution data spherical k-means even comes to the same conclusions as spherical k-principal-components, regardless of starting values. Research that follows up on this thesis can experiment with multiples starting values and different values for k. One might also want to take the running time of the clustering methods into account, as spherical k-means takes 54 seconds for the finance data set, while spherical k-principal components converges after about half the amount of time at 28 seconds. Both runtimes do not include optimisation of k. Obtaining the scree plot for the finance data set using spherical k-means takes just over 2.5 minutes.

When including more variables the results differ on various fronts and different starting values produce different results. PCA and spherical k-means differ for any number of variables, but future

research can be conducted so that the number of clusters and number of components are better aligned with each other.

Finally we further inspect why results obtained by spherical k-principal-components for many variables differ when using different starting values. We argue that k-means starting values are not reliable, because the results are very distinct from the results obtained by spherical k-means and PCA. These differences could have two explanations. First, certain starting values can result in the algorithm ending up in a local optimum. One could think that using starting values close to the optimal values ensure convergence to the global optimum, but this is not true. A simple counterexample is shown in Figure 17.

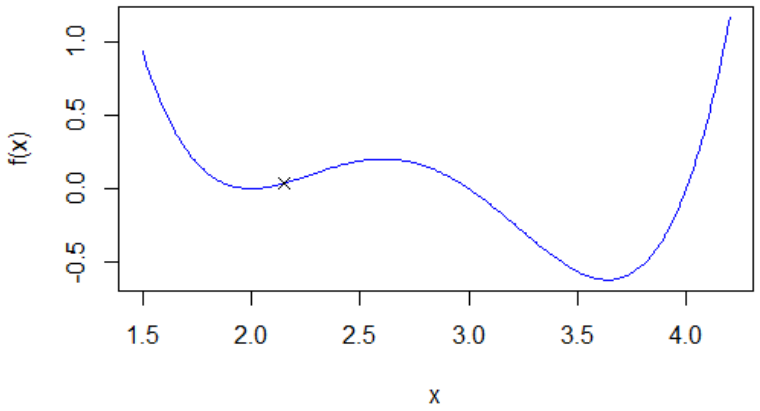


Figure 17: Plot of function $f(x) = (x-2)^2(x-3)(x-4)$.

In the plot, spherical k-means starting values are indicated by the black cross at the coordinate $(x, f(x)) = (2.15, f(2.15) \approx (2.15, 0.035)$. The difference in the x-values between this point and the global optimum is about 1.5, which is less than the distance to the nearest local optimum. Now the random uniform starting values are any point from $x = 5.5$ onwards (i.e. points with a distance greater than 1.5 from the global optimum). In this example the random uniform starting values are further away from the global optimum than the spherical k-means starting values. Yet, the random uniform starting values are likely to end up in the global optimum while the spherical k-means starting values end up in a local optimum.

The second possible explanation for the different results is that spherical k-means starting values actually produce better results. Fomichov & Ivanovs state that spherical k-principal-components work better for concomitant extremes. This should ensure a sparse model, which for this algorithm means we would have a small number of clusters. This explains why many clusters are dominated by one variable, because these clusters would be merged for lower values of k. Decreasing the number of clusters ensures that each cluster has more observations so that clusters are less likely

to be dominated by one variable. This is another argument for why researchers should experiment with different values of k in the future. In short, we have not found any hard evidence to show that spherical k-principal-components with spherical k-means starting values produce incorrect result, but the clusters obtained for the finance data set do seem to point in that direction. To ensure reliable results one should use a set of different starting values and try to find similarities between all the obtained results. If multiple starting values produce the same results, the corresponding conclusion that can be drawn is likely to be true.

References

Bholowalia, P. & Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, *105*(9).

Brodin, E. & Klüppelberg, C. (2014). Extreme value theory in finance. *Wiley StatsRef: Statistics Reference Online*.

Buchta, C., Kober, M., Feinerer, I. & Hornik, K. (2012). Spherical k-means clustering. *Journal of statistical software*, *50*(10), 1–22.

Coles, S., Bawa, J., Trenner, L. & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208). Springer.

Cooley, D. & Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, *106*(3), 587–604.

Dhillon, I. S. & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, *42*(1), 143–175.

Ding, C. & He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on machine learning* (p. 29).

Drees, H. & Sabourin, A. (2019). Principal component analysis for multivariate extremes. *arXiv preprint arXiv:1906.11043*.

Easterling, D. R., Evans, J. L., Groisman, P. Y., Karl, T. R., Kunkel, K. E. & Ambenje, P. (2000). Observed variability and trends in extreme climate events: a brief review. *Bulletin of the American Meteorological Society*, *81*(3), 417–426.

Fomichov, V. & Ivanovs, J. (2020). Detection of groups of concomitant extremes using clustering. *arXiv preprint arXiv:2010.12372*.

From, A. M., Leibson, C. L., Bursi, F., Redfield, M. M., Weston, S. A., Jacobsen, S. J., ... Roger, V. L. (2006). Diabetes in heart failure: prevalence and impact on outcome in the population. *The American journal of medicine*, *119*(7), 591–599.

Heij, C., de Boer, P., Franses, P. H., Kloek, T. & van Dijk, H. K. (2004). *Econometric methods with applications in business and economics*. Oxford University Press.

Hosking, J. R. (1985). Algorithm as 215: Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *34*(3), 301–310.

Janßen, A. & Wan, P. (2020). k-means clustering of extremes. *Electronic Journal of Statistics*, *14*(1), 1211–1233.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, *81*(348), 158–171.

Johnson, R. A. & Wichern, D. W. (2014). *Applied multivariate statistical analysis* (Vol. 5) (No. 8). Pearson Education Limited.

Kanyongo, G. Y. (2005). Determining the correct number of components to extract from a principal components analysis: A monte carlo study of the accuracy of the scree plot. *Journal of modern applied statistical methods*, *4*(1), 13.

Kottas, A. & Sansó, B. (2007). Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, *137*(10), 3151–3163.

Larsson, M. & Resnick, S. I. (2012). Extremal dependence measure and extremogram: the regularly varying case. *Extremes*, *15*(2), 231–256.

Likas, A., Vlassis, N. & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451–461.

Mannering, F. L., Shankar, V. & Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic methods in accident research*, *11*, 1–16.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Pham, D. T., Dimov, S. S. & Nguyen, C. D. (2005). Selection of k in k-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103–119.

Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.

Tippett, L. H. (1925). On the extreme individuals and the range of samples taken from a normal population. *Biometrika*, 364–387.

Valle, S., Li, W. & Qin, S. J. (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research*, *38*(11), 4389–4401.

Yang, R. & Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 1195–1211.

Zhang, X., Zwiers, F. W. & Li, G. (2004). Monte carlo experiments on the detection of trends in extreme values. *Journal of Climate*, *17*(10), 1945–1952.