Erasmus University Rotterdam
Erasmus School of Economics

Bachelor Thesis Econometrics & Operations Research

Spherical k-means on multivariate moderate dimensional data

Author:
Gerlise Chan
471044

Supervisor:
Dr. Phyllis Wan
Second assessor:
Dr. Alex Koning

Abstract

Extremal events like the financial crisis in 2008 and the COVID-19 pandemic in 2020 have shown how important accurate assessment of these rare events are. Extreme Value Theory (EVT) is a widely used approach to quantify the risks in different fields. It is often difficult to find suitable parametric models for a variety of datasets in the multivariate setting. This paper focuses on an exploratory procedure that gives an overview of the extremal and non-extremal dependence structure. In particular, three different datasets from different fields are used to investigate the extremal and non-extremal data. The procedure uses multivariate extreme value analysis and spherical k-means to estimate the spectral measure that reveals the dependence structure. The results of the extremal and non-extremal data are analyzed and compared. This research concludes that this method provides relevant patterns and is able to classify extremal events which gives different results when applied on non-extremal data.

July 4, 2021

Contents

1 Introduction

The financial crisis in 2008, the most severe global recession after the Great Depression. The crisis lead to the bursting of the US housing bubble and severe damage to financial institutions like the bankruptcy of the Lehman Brothers. How did it all came down? According to Salmon (2012), it was the ignored limitations of the Gaussian copula function used on CDOs by Li (2000). It was a formula that was able to model complex risks in an easy way. It was adopted by many as it was making them money fast. However, it was a fatally flawed way to assess risks.

Another approach to quantify risks is using Extreme Value Theory (EVT). The crucial element of EVT is the extrapolation, estimating the small probabilities of rare events or events that have yet to be observed using observations from datasets that show extremal behavior. One extreme event may not result in a financial crisis, but when several extreme events occur simultaneously it can be catastrophic. For example when extreme rainfall occurs simultaneously in an area it increases the risk of flooding (Thibaud et al. (2013)). If extremal dependence is not correctly specified, the severity of risk can be under estimated.

While EVT in the univariate setting is well established, the application in the multivariate setting is still limited. Due to the complexity, most parametric models are found in the bivariate or low-dimensional settings. In addition, it is often difficult to find a suitable parametric model for a variety of datasets. Therefore it is useful to find an exploratory procedure to give an overview of the important extremal dependence structures before fitting a parametric model. Therefore the following research question is proposed: "*How does spherical k-means find patterns and classify rare events on moderate dimensional data and how does the dependence structure of extremal data differ from the dependence structure of non-extremal data?*"

To answer this research question, the spherical k-means algorithm in combination with extremes is discussed. To investigate if similar patterns in the dependence structure exists between the extremal data and non-extremal data, the results are compared. This paper uses three different datasets from different fields to investigate the dependence structure of the cluster centers. The first dataset contains 30 industry portfolios compiled and posted by the Kenneth French Data Library between 1950-2015. To investigate the losses, the data is multiplied by -1. The second dataset contains nutrients information from the food and beverage consumption consumed 24 hours prior to the interview by participant in the interview "What We Eat in America" from 2015-2016 NHANES report. The third dataset contains the electricity consumption of households categorized by the 57 ACORN types between 2007-2010 provided by AECOM Building Engineering (2018). The ACORN system provides information about the financial and living situation of each household.

Spherical k-means shows that cluster centers can reveal and explain the dependence structures. The results from the financial portfolio losses suggests that there is a difference between the dependence

structure using extremal data and using non-extremal data. The non-extremal data scarcely displays losses that occurred by multiple industries simultaneously. Therefore the non-extremal data does not capture the dependence structure of extremes. The results from the dietary intake data suggests that high intake of some nutrients are accompanied by other nutrients. From the non-extremal data, most of the nutrients were strongly independent. The only significant clusters that appeared in both extremal and non-extremal data is the cluster vitamin K and lutein. These nutrients occurred simultaneously in high doses. The results from the electricity consumption suggests that the dependence structure of non-extremes can not capture the dependence structure of extremes. The clusters that are formed share no similarity. Looking at the consumption pattern, extreme electricity usage and large electricity consumption is observed during cold spells, heavy snowfall, extreme rainfall and above average temperatures.

In section 2, the existing research and their main findings will be discussed. In section 3, a description of the three datasets and adjustments will be given and discussed. In section 4, the procedure and the methods will be explained. In section 5, the results of the procedure on the three dataset will be shown. Finally in section 6 the main findings, limitations of our research and further research recommendations will be given.

2 Literature

Extreme value analysis is a statistical method to analyze the stochastic behavior of extremal observations. Extremal observations in a dataset contain informative signals in the distribution tails. A crucial element in extreme value analysis is estimating the probabilities of events that have not been observed, this is called extrapolation. Therefore we are interested in the limiting probability distributions tails. The central limit theorem (CLT) is often associated with the limiting probability distribution. Extreme value theory (EVT) is therefore similar to the CLT. While CLT concerns about the behavior in the entire distribution, EVT focuses on the behavior in the tails.

In general there are two approaches to extract the extremal observations: the block maxima and the peak-over-threshold. In this paper we will be focusing on the peak-over-threshold approach. Both approaches uses a limiting probability distribution to characterize the data. The block maxima approach uses a Generalized Extreme Value distribution (Fisher and Tippett (1928)) and the peak-over-threshold approach uses a Generalized Pareto distribution (Pickands III et al. (1975) and Balkema and De Haan (1974)).

The application of extreme value analysis on the univariate setting is well researched, but the methods for the multivariate setting are still limited. Especially in higher dimensions, as data is getting higher dimensional, the extremal dependence of a random vector reveals complicated structures (Coles et al. (1999)). Therefore it is crucial to find tools that can reduce the dimension.

Engelke and Ivanovs (2020) has grouped different approaches in three areas of research. The first area of research are the methods from unsupervised learning. For example, Principal Component Analysis (PCA) and clustering. These dimension reduction techniques are used to visualize extremal dependence and are used for exploratory analysis. This paper concentrates on this area of research.

Janßen and Wan (2020) have shown a new procedure for analyzing extremal dependence of random vectors. They applied the clustering algorithm spherical k-means to estimate the spectral measure. It allows for a more gradual view on dependencies instead of classifying groups of random variables which shared the same dependencies. In their data examples, they have shown that the dependence structures are revealed by "extremal prototypes" in the largest observations. Their method provided insights on datasets with moderate dimensions.

Similarly using a clustering approach, Chautru (2015) proposed a clustering approach in combination with spectral measure analysis to reduce the dimension. Using Principal Nested Spheres, an algorithm that adapts PCA to Riemannian manifolds by Jung et al. (2012), observation angles are projected on a space that has lower dimension. Using spherical k-means, groups of interest are identified. Numerical experiments and a case study using a dietary risk assessment dataset are conducted to analyze extreme dependencies. One advantage of the approach is that it can incorporate different types of extreme dependencies.

Cooley and Thibaud (2019) used related covariance matrix decomposition techniques to describe high-dimensional tail dependence. In non-extreme settings, the covariance matrix is often used to describe dependency between data. For example, PCA is an eigendecomposition of the covariance matrix. They developed tools to perform exploratory analysis of extremal dependence in high dimensions.

The second area of research mentioned by Engelke and Ivanovs (2020) is the rare event analysis. It looks at which sub-groups of variables are likely to take large values simultaneously. Goix et al. (2017) designed a method DAMEX (Detecting Anomalies with Multivariate EXtremes) that reveals the sparsity pattern in the extremal dependence structure while obtaining bounds for the accuracy of the estimation procedure. It is used as a preprocessing step to reduce the high dimensions of data. They show that this algorithm is suitable for real world large-scale learning problems and show that the performance is better than standard anomaly detection approaches. To encounter the ambiguous structure issue with DAMEX, Chiapino and Sabourin (2016) came up with the CLEF (CLustering Extreme Features) algorithm. This algorithm groups together features that take extreme values simultaneously. It makes use of the Apriori algorithm which reduces the number of subsets to bypass computational issues. The results of the simulation and real dataset show that meaningful information of the dependence structure of extremes is retrieved.

The third area of research mentioned by Engelke and Ivanovs (2020) are the conditional independence

structures and graphical models. These allow for higher dimensional distributions to be decomposed into low-dimensional components. Gissibl, Klüppelberg, et al. (2018) considered a new recursive structural equation model where the model is max-linear in terms of the noise variables. The model structure is represented by a directed acyclic graph. They found that the minimum directed acyclic graph represents the recursive structural equations of the variables which yields the representations of the vector components. Engelke and Volgushev (2020) show that the underlying graph of the extremal graphical models conceals conditional independence structure and is able to visualize the complex extremal dependence structure. They show a new summary statistic, extremal variogram, that is able to consistently recover the true underlying tree by using the statistic as weights for the minimum spanning tree.

3 Data

3.1 Financial portfolio losses

The first dataset is the daily returns of 30 industry portfolios available at the Kenneth French Data Library. The considered data is the value-averaged daily returns between 1950-2015, this results in 16694 observations. The same dataset is analyzed in Cooley and Thibaud (2019), where a method related to PCA is used to describe extremal dependence. Janßen and Wan (2020) used a clustering approach for this dataset which is applied in our paper.

The 30 industry portfolios were constructed by assigning each NYSE, AMEX and NASDAQ stock to an industry portfolio based on its four digit SIC code. When the SIC codes are not available, CRSP SIC codes are used. All returns are multiplied by -1 as we are interested in the portfolio losses. A description on the industry definitions can be found at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_30_ind_port.html. Some relevant statistics of the dataset are given in Table 1.

Table 1: Relevant statistics for the financial portfolio losses.

Industry	Mean	SD	Industry	Mean	SD	Industry	Mean	SD
Food	-0.0499	0.8362	Cnstr	-0.0458	1.1443	Telcm	-0.0435	1.0189
Beer	-0.0520	1.0641	Steel	-0.0378	1.4583	Servs	-0.0524	1.2198
Smoke	-0.0633	1.2727	FabPr	-0.0456	1.1681	BusEq	-0.0534	1.4022
Games	-0.0512	1.419	ElcEq	-0.0565	1.2937	Paper	-0.0475	1.0009
Books	-0.0448	1.099	Autos	-0.0459	1.3426	Trans	-0.0467	1.1466
Hshld	-0.0476	1.0232	Carry	-0.0552	1.2263	Whlsl	-0.0460	1.0132
Clths	-0.0453	1.1215	Mines	-0.0397	1.4698	Rtail	-0.0501	1.0421
Hlth	-0.0546	1.0259	Coal	-0.0500	2.0435	Meals	-0.0549	1.1924
Chems	-0.0461	1.1344	Oil	-0.0525	1.2165	Fin	-0.0481	1.1292
Txtls	-0.0458	1.2155	Util	-0.0419	0.7704	Other	-0.0387	1.0809

3.2 Dietary intake

The second dataset is the food and beverage consumption consumed by participants in the interview "What We Eat in America" from the 2015-2016 NHANES report. It is a nationally representative sample where Asians, Hispanics, blacks, low-income whites/others and whites/others 80+ years are oversampled. The five-step USDA Automated Multiple-Pass Method is used for collecting interviewer-administered dietary recalls. The data is recorded 24 hours prior to the interview and contains nutrients information calculated from the observations. The dataset is available at https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DR1TOT_I.XPT. The same procedure in Janßen and Wan (2020) is applied. Chautru (2015) conducted a similar analysis for a smaller dataset.

The dataset contains 65 nutrients information. To investigate if high doses of some nutrients have a negative health effect, only 38 nutrients are considered. Removing missing values resulted in 8327 observations. Some relevant statistics of the nutrients information are given in Table 2.

Table 2: Relevant statistics for the dietary intake.

Nutrient	Mean	SD	Nutrient	Mean	SD	Nutrient	Mean	SD
Calories	1947.6383	928.9864	Alpha_Car	342.0905	1117.5436	VitaminC	77.7471	83.0852
Protein	73.3603	40.7331	Beta_Car	1828.2271	3841.0411	VitaminD	4.9617	5.3604
Carbs	237.6535	117.0532	Beta_Crypt	86.3618	236.4084	VitaminK	96.0572	159.2094
Sugar	103.5942	66.6708	Lycopene	4428.6874	7783.5085	Calcium	913.2091	569.7381
Fiber	15.463	10.1331	Lutein	1279.0606	3233.4637	Phosphor	1265.4856	650.5701
Fat	76.2669	44.0167	VitaminB1	1.4897	0.8525	Magnesium	263.0261	142.8278
Sat_fat	25.4493	16.0396	VitaminB2	1.9168	1.1552	Iron	13.5543	8.1519
Mono_unsat_fat	26.5084	16.1554	Niacin	22.9517	14.5548	Zinc	10.1882	6.5487
Poly_unsat_fat	17.3837	11.8975	VitaminB6	1.8502	1.5579	Copper	1.0576	0.6801
Cholesterol	270.2649	234.2611	Folate	363.8087	241.6397	Sodium	3198.432	1792.0135
VitaminE	7.9468	6.0296	Folic_acid	174.2404	177.5652	Potassium	2332.1382	1159.7081
Retinol	412.1855	416.9526	Choline	298.9528	189.5002	Selenium	104.358	63.0756
VitaminA	581.7371	553.1031	VitaminB12	4.5007	4.3216			

3.3 Electricity consumption

The third dataset is the average electricity consumption during peak hours categorized by all ACORN types in Great Britain. The dataset was executed between 2007-2010 by the Energy Demand Research Project. The dataset measured the electricity consumption of over 60000 households. Each households is categorized using the ACORN system. The 57 ACORN types provides information about the living and financial situation of each household. The peak hours are given by 16:30-19:30. The dataset is provided by AECOM Building Engineering (2018), where a detailed description of the dataset is given.

The dataset contains half-hourly electricity consumption measured in units of kilowatt-hour. The considered data is obtained by taking the average of the electricity consumption during peak hours of the

categorized households with a certain ACORN type. This resulted in 861 observations. The description of the ACORN types are given in Table 4 in the Appendix and some relevant statistics of the electricity consumption per ACORN type are given in Table 3.

Table 3: Relevant statistics for electricity consumption.

Type	Mean	SD	Type	Mean	SD	Type	Mean	SD
1	0.5324	0.1358	20	0.2946	0.0792	39	0.3686	0.0797
2	0.4836	0.1174	21	0.3558	0.0719	40	0.3482	0.0807
3	0.4536	0.1243	22	0.3077	0.0886	41	0.3294	0.0793
4	0.4617	0.1175	23	0.4279	0.1269	42	0.3061	0.0729
5	0.4344	0.1122	24	0.3241	0.0766	43	0.2689	0.0629
6	0.4368	0.1398	25	0.3441	0.0772	44	0.3490	0.0811
7	0.4124	0.1088	26	0.3908	0.0898	45	0.2954	0.0717
8	0.3614	0.0915	27	0.4010	0.0982	46	0.3198	0.0770
9	0.4226	0.1073	28	0.3901	0.0898	47	0.3294	0.0744
10	0.4437	0.1042	29	0.3803	0.0899	48	0.2921	0.0712
11	0.4284	0.1188	30	0.3527	0.0869	49	0.3618	0.0818
12	0.4712	0.1935	31	0.3780	0.0965	50	0.2270	0.0527
13	0.4177	0.1028	32	0.2842	0.0813	51	0.2622	0.0698
14	0.3762	0.0937	33	0.3404	0.0848	52	0.2792	0.0707
15	0.5132	0.1549	34	0.3406	0.0898	53	0.2111	0.0499
16	0.3659	0.1123	35	0.2033	0.0576	54	0.2330	0.0567
17	0.2675	0.0754	36	0.3593	0.0977	55	0.2612	0.0630
18	0.4960	0.1930	37	0.3495	0.1288	56	0.2743	0.0790
19	0.3782	0.0988	38	0.2675	0.0662	57	0.4828	0.1346

4 Methodology

In this section, a brief overview of the multivariate extreme value theory is given. Afterwards, a summary of the spherical k-means algorithm is discussed. At last, the procedure that will be applied on the three datasets is described.

4.1 Multivariate Extreme Value Theory

To study the small probabilities of rare events, extreme value analysis is used to provide accurate assessment. Extreme values are observations in the distribution tails that contain informative signals, therefore we are interested in the limiting probability models. Let a d-dimensional vector $\mathbf{X} = (X_1, ..., X_d)$ and $(X_1^i, ..., X_d^i)$ be the i.i.d. copies of $\mathbf{X}$ where $1 \leq i \leq n$. The maximum $M_j = \max\{X_j^1, ..., X_j^n\}$ where $1 \leq j \leq d$. According to the Fisher-Tippett theorem by Fisher and Tippett (1928), if sequences of constants $a_j^n > 0$ and $b_j^n \in \mathbb{R}$ exists where $n \to \infty$ then the limit of the probability converges to a max-stable

nondegenerate distribution function G.

$$P\Big(\frac{M_1 - b_1^n}{a_1^n} \leq x_1, ..., \frac{M_d - b_d^n}{a_d^n} \leq x_d\Big) \xrightarrow{d} G(x_1, ..., x_d) \tag{1}$$

To model the extreme values, the peak-over-threshold approach is used. A threshold is used to divide the extremes from the data. Let u be a threshold and $F(x_j)$ the distribution function, then the excess distribution is as follows where $\bar{F}(u)$ is the exceedance probability:

$$F_u(x_j) = P(X_j - u \leq x_j | X_j > u) = (F(u + x_j) - F(u))/\bar{F}(u), \qquad 0 \leq x_j < \infty \tag{2}$$

Using Pickands–Balkema–de Haan theorem by Pickands III et al. (1975) and Balkema and De Haan (1974), $F_u(x_j)$ can be approximated using the nondegenerate distribution G which is a Generalized Pareto Distribution (GPD). The marginals in (2) are normalized using the following equation, this transforms the marginals to the same scale.

$$\mathbf{Y} = \Big(\frac{1}{1 - F_1(X_1)}, ..., \frac{1}{1 - F_d(X_d)}\Big) \tag{3}$$

The convergence in (1) holds if and only if (3) satisfies the following where S is a probability measure that can be obtained from the limiting behavior and B is any S-continuity-Borel-set.

$$\lim_{u \to \infty} P\Big(\frac{\mathbf{Y}}{\|\mathbf{Y}\|} \in B \;\Big|\; \|\mathbf{Y}\| > u\Big) = S(B) \tag{4}$$

The spectral measure S describes the dependence structure of $\mathbf{Y}$ uniquely and it explains the angle of the extremal observations. This is the measure of interest that is estimated using spherical k-means.

4.2 Spherical *k*-means

The k-means algorithm is a well known and widely used unsupervised partitional clustering approach. The standard k-means algorithm assigns observations to the closest cluster center. The following general objective is minimized given the probability measure P on $\mathbb{B}(\mathbb{R}^d)$ where $\mathbb{B}(.)$ is the Borel σ-algebra, the number of clusters k and the set of cluster centers $A = \mathbf{a}_1, ..., \mathbf{a}_k$ where $\mathbf{a}_i \in \mathbb{R}^d$ for $i = 1, ..., k$ and $k \in \mathbb{N}$.

$$W(A, P) := \int_{\mathbb{R}^d} min_{\mathbf{a} \in A} d(\mathbf{x}, \mathbf{a}) P(d\mathbf{x}) \in [0, \infty) \tag{5}$$

For standard k-means the Euclidean distance is used as distance function or dissimilarity function. For spherical k-means, the distance is measured using the angle between two points. The angular dissimilarity measure, the cosine dissimilarity, is defined by Dhillon and Modha (2001) as follows:

$$d(\mathbf{x}, \mathbf{a}) = 1 - cos(\mathbf{x}, \mathbf{a}) = 1 - \frac{\langle \mathbf{x}, \mathbf{a} \rangle}{\|\mathbf{x}\| \|\mathbf{a}\|} \tag{6}$$

The optimal number of clusters k is chosen using "elbow plot" where the minimized mean distance $W(A, P)$ in (5) is plotted against multiple k values. The optimal k is found when a significant decrease of the criterion against k is observed, the "elbow". To interpret the cluster centers and investigate the dependence structures, heat maps are used.

4.3 Procedure

To estimate the measure of interest S to explains the angle or the dependence structure of the extremal distribution, the procedure follows the following steps:

1. The function *ecdf* in R is used to compute the empirical cdf distribution of each dimension for each dataset. Using the Fréchet transformation described in (3), each dimension is transformed such that all marginals are on the same scale.

2a. For the extremal observations, a fraction of the transformed data with the largest Euclidean norm is kept by using a threshold.

2b. For the non-extremal observations, the extremal observations in 2a. are omitted. To avoid observations contributing to noise, another threshold is used to keep a fraction of the transformed data.

3. The obtained subset in 2a. or 2b. is then normalized by projecting the data onto the unit sphere.

4. Spherical k-means is applied to the projected observations using the R package "skmeans" by Buchta et al. (2012) with *method* pclust, $nruns = 1000$ and $maxchains = 100$.

The procedure is from Janßen and Wan (2020). They showed a theorem that contribute to the statistical inference of S. Under suitable convergence of the estimated S, the corresponding empirical cluster centers of the estimated spectral measure converge to their theoretical counterparts when the estimated spectral measure converges.

5 Results

In this section, we discuss the dependence structure of the extremal and non-extremal data. For each dataset, we start by analyzing the results of the extremes. Thereafter the results of the non-extremes are analyzed.

5.1 Financial portfolio losses

The procedure explained in the methodology is applied. First the marginals are computed and transformed as described in (3). Using the Euclidean norm, 5% of the largest observations are selected. This subset of the data is then normalized by projecting the data onto unit sphere. Afterwards spherical k-means is applied. The number of clusters k is determined using a classic approach, "elbow plot". In Figure 1 the optimal k can not be determined, therefore the results of $k = 5$ and $k = 10$ are analyzed and compared.

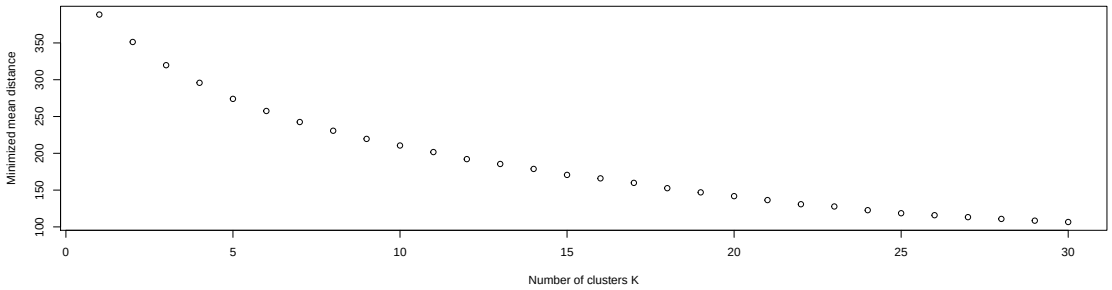


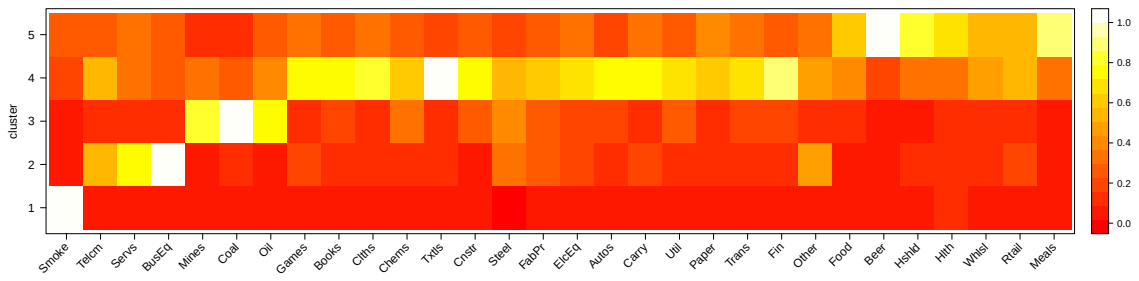
Figure 1: Elbow plot of the financial portfolio losses using extremal data.

Using the following equation, the cluster centers are re-normalized such that the component with the highest value is scaled to 1. Each row in the heat map in Figure 2 represents a cluster center. Using (7), the box with the lightest color corresponds to the component with the largest value relative to other components in this row.

$$\mathbf{a}_i = (a_1^i, ..., a_d^i) \rightarrow \Big(\frac{a_1^i}{\max\limits_{h=1,..,d} \{a_h^i\}}, ..., \frac{a_d^i}{\max\limits_{h=1,..,d} \{a_h^i\}} \Big) \tag{7}$$

It is observed that the clusters separate the industries into different sectors. For Cluster 1, it is observed that the tobacco industry is independent to all other industries. Cluster 2 focuses on the business and IT related industries, Cluster 3 focuses on the energy and material industries, Cluster 5 focuses on industries that are consumer oriented and Cluster 4 focuses on the rest.

The plot in Figure 2 shows the time period of the extreme losses. It is observed for Cluster 1 and Cluster 2 that most extreme losses are made in the period around 2000. This is when the tobacco industry was suffering from massive lawsuits and the bursting of the dot-com bubble which caused bankruptcies in internet and tech businesses. One of the causes of extreme losses in Cluster 3 is the turn of the millennium and the rise of alternative sources of energy. The extreme losses in Cluster 5 are due to the US recessions and oil crisis in 1973. Consumer goods were heavily affected by these events and a combination of the dot-com bubble and financial crisis explain the extreme losses in Cluster 4.



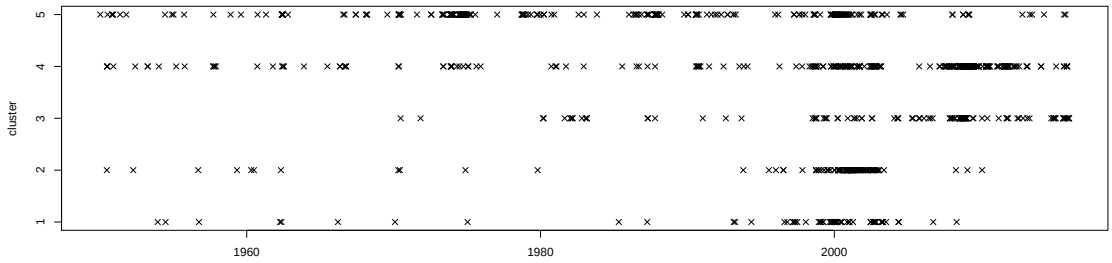


Figure 2: Clustering results of the financial portfolio losses using extremal data for $k = 5$.

The results in Figure 3 can be interpreted in a similar way. Most cluster centers contain one or a few large components which show signs of strong independence. Some results remain from the analysis of $k = 5$. So is observed that the time period of extreme losses using $k = 10$ show similar patterns as the analysis of $k = 5$. The extreme losses due to the oil crisis and financial crisis (dot-com bubble) are reflected in the plot. In addition, some clusters focuses on the same industries as explained in $k = 5$. The industries are: the consumer oriented (Cluster 4), the business and IT related (Cluster 6), the energy and material sectors (Cluster 5) and the industries linked to the financial sector (Cluster 3).

As a long horizon is used on the time series, it can not be assumed that stationarity holds. The dependence structure can therefore change over time. In addition, the observations are not i.i.d. distributed which was an assumption in Section 4.1, therefore the estimates are biased.

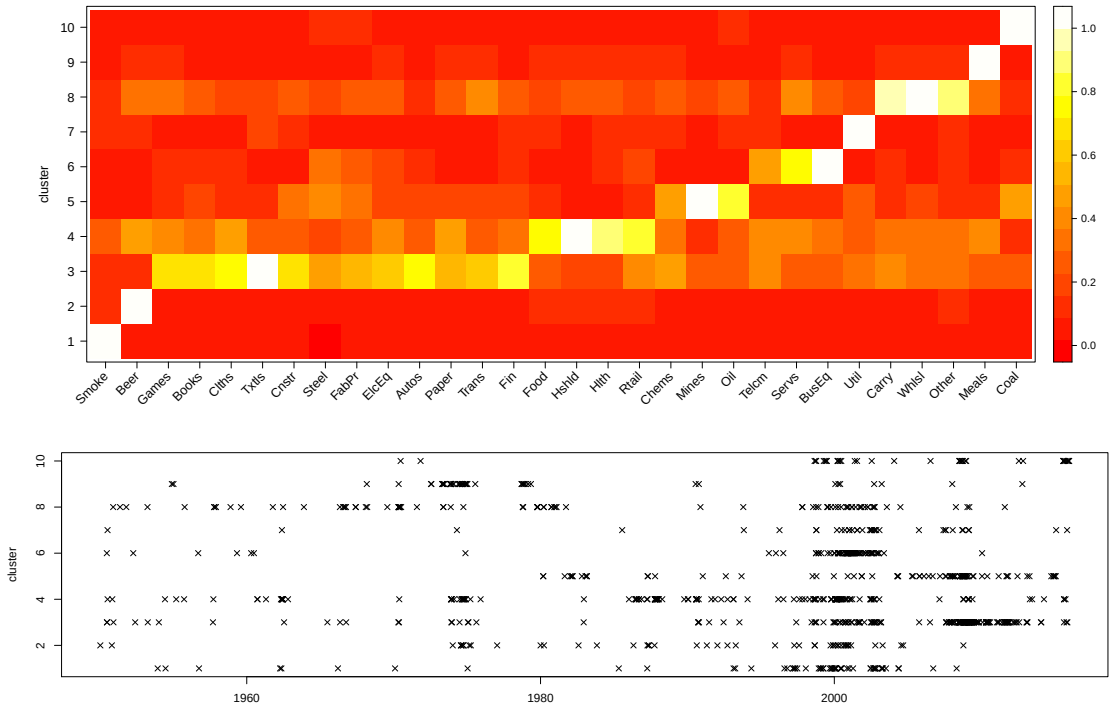


Figure 3: Clustering results of the financial portfolio losses using extremal data for $k = 10$.

The extremal observations are removed to analyze the non-extremal data. From the residual data only

20% of the largest Euclidean norm observations are selected to avoid contribution to noise. The "elbow plot" in Figure 14 is used to select the optimal number of clusters. The analysis of $k = 10$ and $k = 15$ are compared as the optimal k can not be determined.

From the heat map in Figure 4 is observed that some industries are clustered together. Cluster 6 contains many large components due to normalization of the cluster centers using (7). This means that multiple industries are jointly dependent and cause losses simultaneously. Some significant cluster are the textile and apparel industry (Cluster 5) and the consumer oriented and autonomous transport industry (Cluster 6). Most losses in the plot can be explained by the oil crisis and financial crisis.

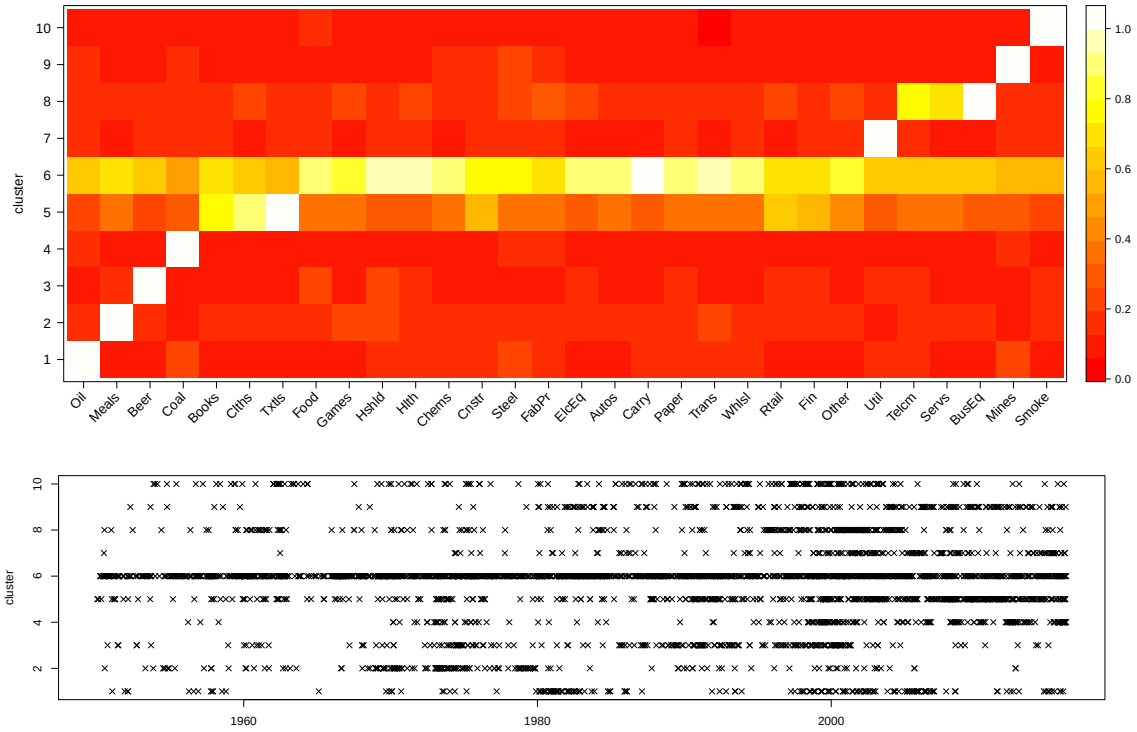


Figure 4: Clustering results of the financial portfolio losses using non-extremal data for $k = 10$.

The analysis of $k = 15$ in Figure 5 can be interpreted in a similar way. Most of the losses are due to the financial crisis and oil crisis. However for each cluster, losses are scattered throughout the years. The clusters that remain from the previous analysis are: the business equipment industry, the textile and apparel industry, the oil industry, the tobacco industry, the utilities sector, the metal industry, the coal industry, the alcoholic beverage industry and the hospitality industry.

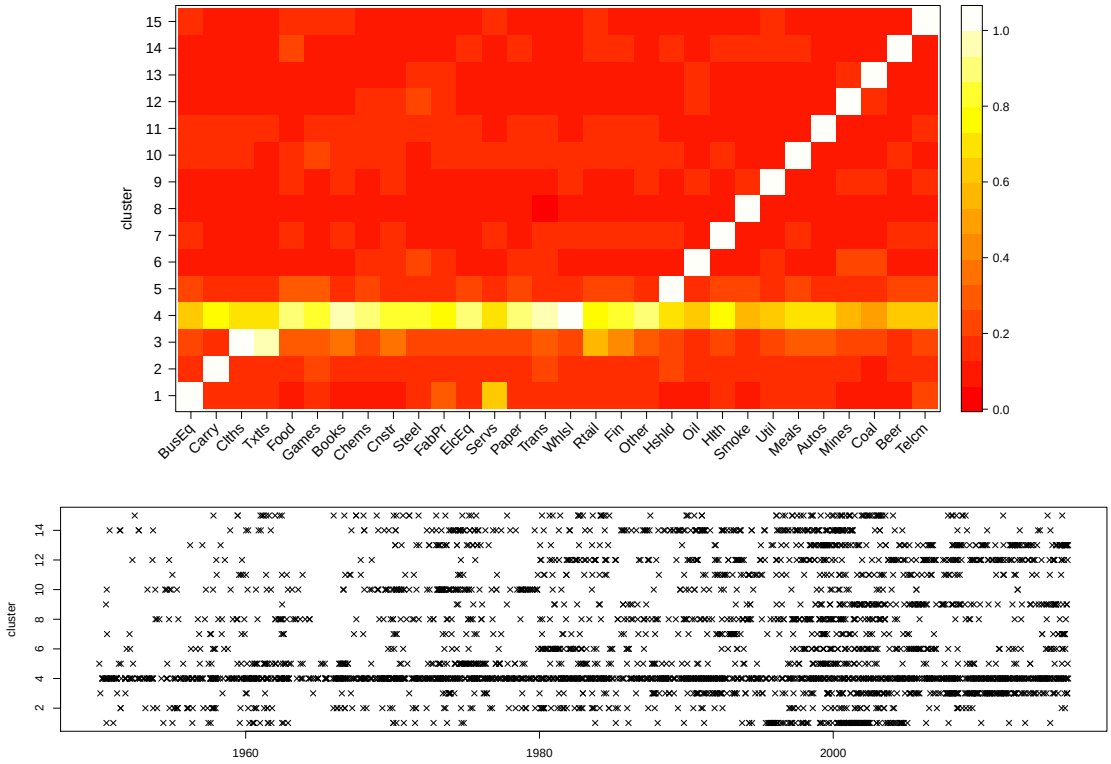


Figure 5: Clustering results of the financial portfolio losses using non-extremal data for $k = 15$.

5.2 Dietary intake

The same procedure as in Section 5.1 is applied on this dataset. Using 5% of the largest Euclidean norm observations, the data is normalized by projecting the observations on the unit sphere. The optimal k is determined using the "elbow plot" in Figure 6. The results of $k = 15$ and $k = 20$ are analyzed and compared as the optimal k can not be determined.

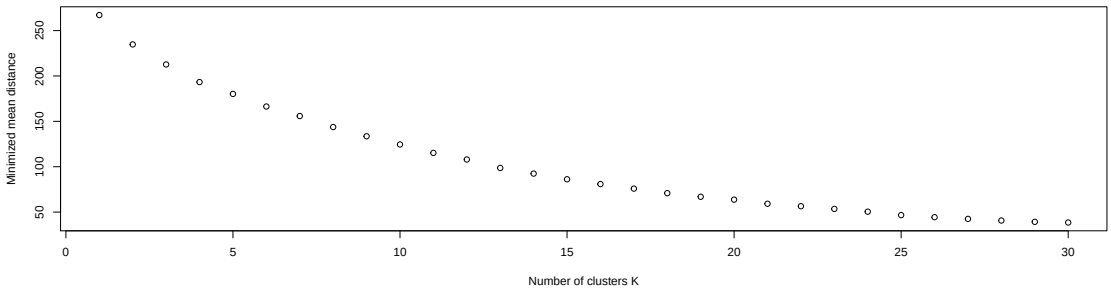


Figure 6: Elbow plot of the dietary intake using extremal data.

The clusters in the heat maps can be interpreted using the same approach in the financial portfolio losses data. The values of each cluster center are normalized, therefore the box with the lightest color corresponds to the component with the largest value relative to other components in this row. It is observed that most clusters separate the nutrients into groups. This indicates the independence of some nutrients. The results from $k = 15$ and $k = 20$ show some similar results. Significant clusters are formed

by vitamin B2, vitamin B6 and niacin, by vitamin K and lutein, by carbs and sugar, by folate and folic acid, by vitamin B1 and iron, and by fat and fatty acids which are accompanied by high intake of calories.

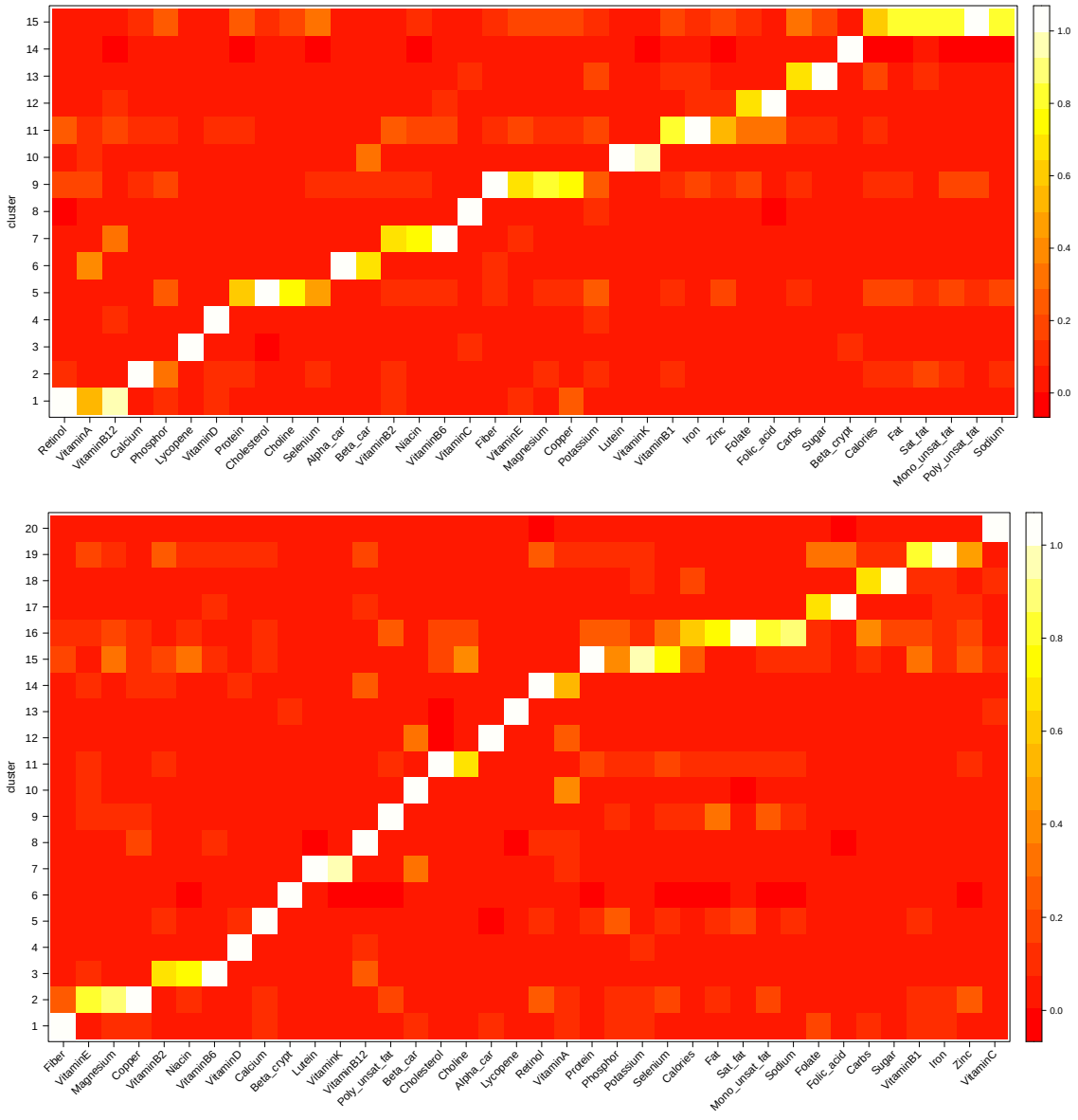


Figure 7: Clustering results of the dietary intake using extremal data for $k = 15$ and $k = 20$.

To analyze the dependence structure of the non-extremal data, the extremal observations are removed. Afterwards 30% of the largest Euclidean norms are selected and clustered. The elbow plot in Figure 15 does not show an optimal k, therefore the analysis of $k = 5$ and $k = 10$ is compared. Significant clusters in both analysis are grouped by cholesterol and choline and by lutein and vitamin K.

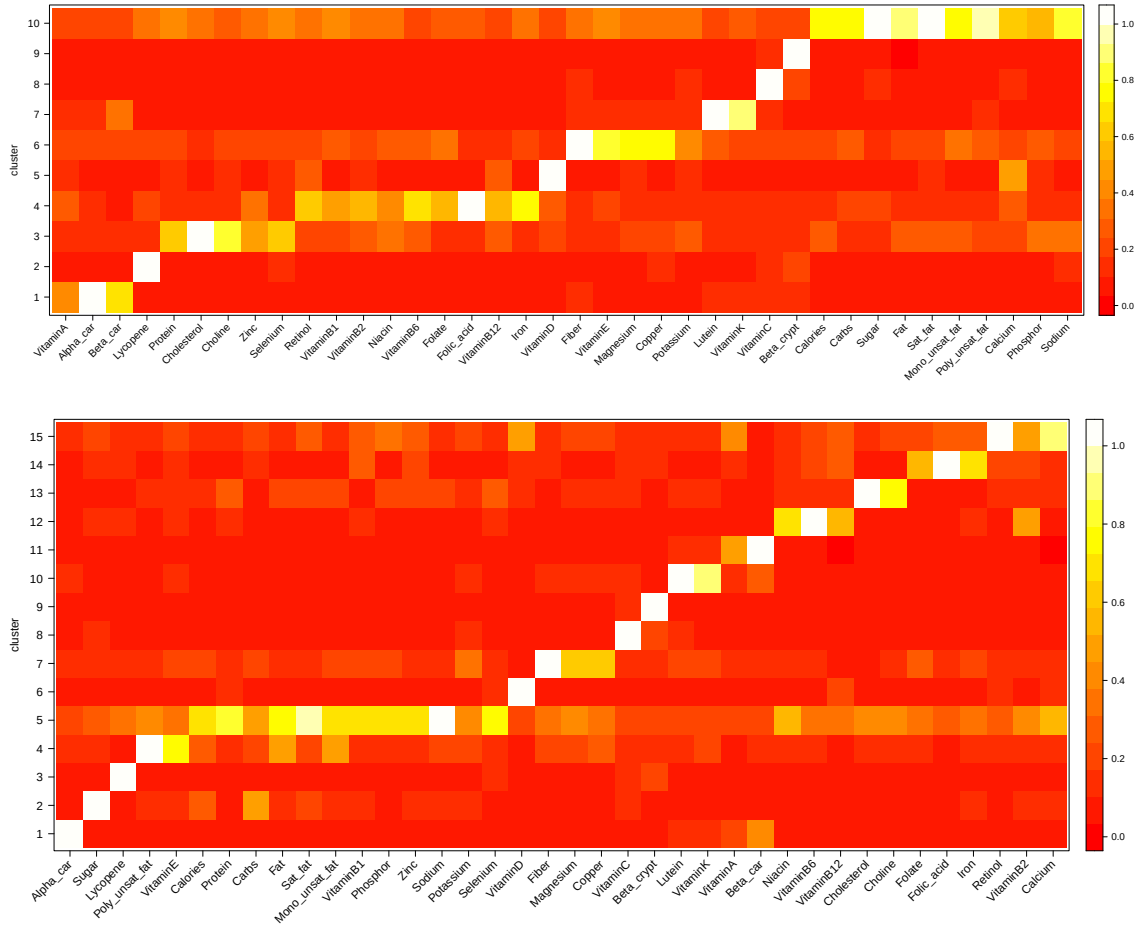


Figure 8: Clustering results of the dietary intake using non-extremal data for $k = 10$ and $k = 15$.

5.3 Electricity consumption

The same procedure as in Section 5.1 is applied on this dataset. Using 10% of the largest Euclidean norm observations, this subset is normalized by projecting the data on the unit sphere. The optimal k is then determined using the "elbow plot" seen in Figure 9. Like the other two datasets, the choice of k can not be determined. Therefore the analysis of $k = 10$ and $k = 15$ is considered.

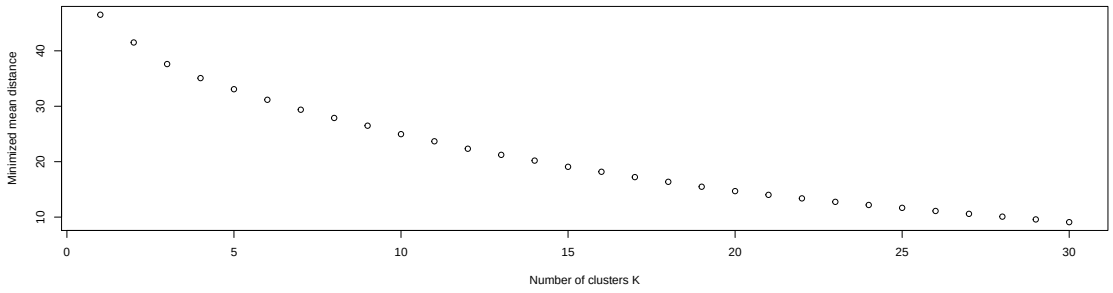


Figure 9: Elbow plot of the electricity consumption using extremal data.

Using heat maps, the dependence structure or the data can be investigated. The cluster centers are

normalized using (7), therefore the lightest box in the row corresponds to the component with the largest value relative to the other components in the same row.

In Figure 10, the ACORN types are grouped together in 10 clusters. Cluster 1 focuses on wealthy well-off households living in larger houses. Cluster 2 focuses on older households living in detached homes. Cluster 3 focuses on larger families and older well-off households living in larger houses or in rural areas. Cluster 4 focuses on older professionals living in suburban houses and apartments. Cluster 5 focuses on young households living in flats. Cluster 6 focuses on home owning middle income households. Cluster 7 focuses on low income households. Cluster 8 focuses on families and single parents living in counsil flats. Cluster 9 focuses on singles and multi-ethnic households living in crowded high rise flats. At last Cluster 10 focuses on mostly communal population.

From the plot in Figure 10 it is observed that Cluster 1 consumed extreme amounts of electricity in the winter of 2009-2010. The winter of 2009-2010 is the coldest winter since 1978-1979 with the coldest winter recorded for northern Scotland. Cluster 2 and Cluster 7 both consumed extreme amounts of electricity in the winter of 2008-2009. The winter of 2008-2009 is known for its cold spells and overall colder average temperatures. Besides consuming extreme amounts of electricity in the winter, extreme consumption is observed in April 2008 by Cluster 3, Cluster 6, Cluster 8 and Cluster 9. This extreme electricity usage can be explained by the overnight extreme snowfall in the UK. Cluster 4 and Cluster 6 consumed extreme amounts of electricity in the summer. So is observed that such extreme event occurred for Cluster 4 in August 2008 and June 2009. For Cluster 6 this occurred in June 2009. Cluster 10 only consumed extreme amounts of electricity in the April 2008. The rainfall in August 2008 was recorded the fourth highest August rainfall since 1962. June 2009 had the 29th highest mean maximum temperature since 1900. The use of air conditioning and heaters in these months explain the extreme electricity usage in these months.

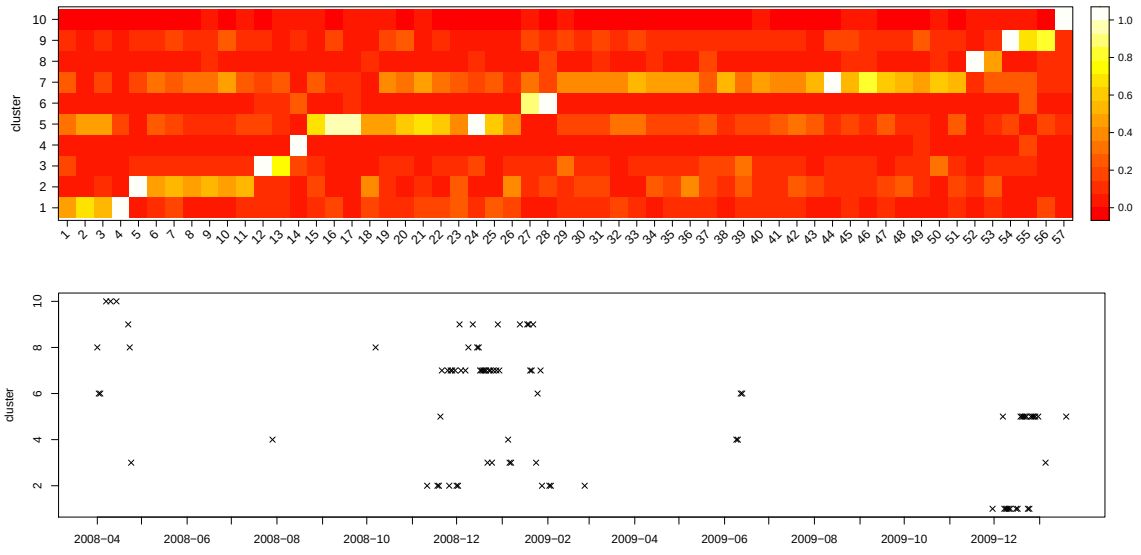


Figure 10: Clustering results of the electricity consumption using extremal data for $k = 10$.

The results in Figure 11 can be interpreted in a similar way. Some results remain from the analysis of k = 10. The clusters that remain from the previous analysis are: the wealthy families with larger houses (Cluster 1), the older households living in detached homes (Cluster 2), young households (Cluster 4) and the communal population (Cluster 15). In addition, the time period of extreme electricity consumption of $k = 15$ show similar patterns. The main events of extreme electricity usage are April 2008, August 2008, winter of 2008-2009, June 2009 and the winter of 2009-2010. The matching groups have a similar consumption pattern.

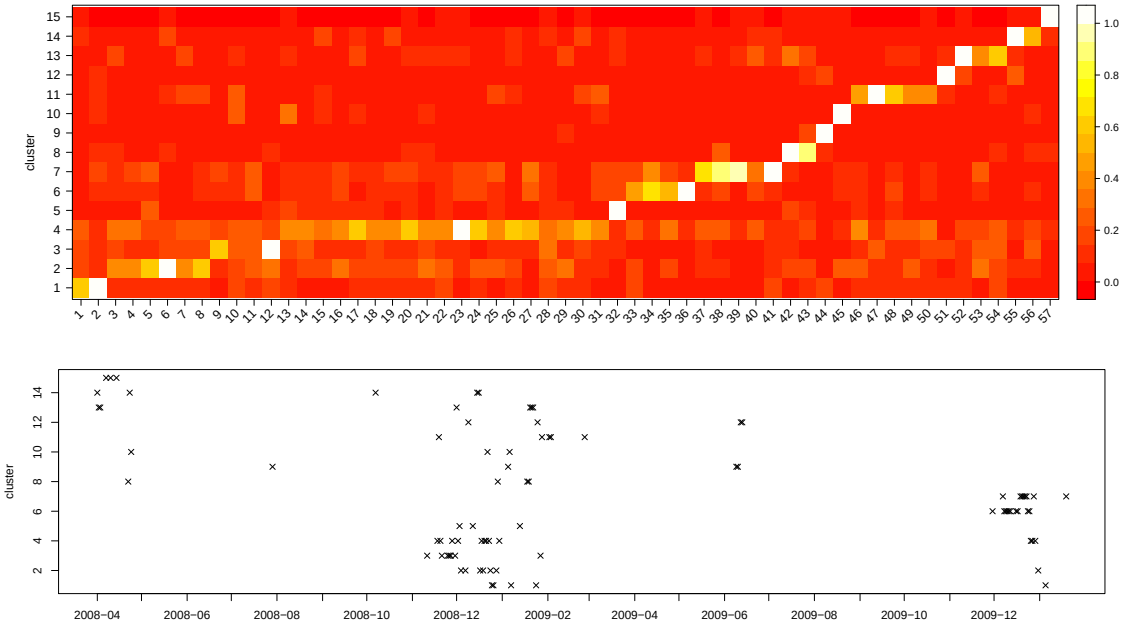


Figure 11: Clustering results of the electricity consumption using extremal data for $k = 15$.

After removing the extremal observations from the normalized and transformed dataset, 15% of the largest Euclidean norm observations are selected to avoid contribution to noise. Applying spherical k-means, Figure 16 shows that the optimal k can not be determined. Therefore the analysis of $k = 5$ and $k = 10$ is considered.

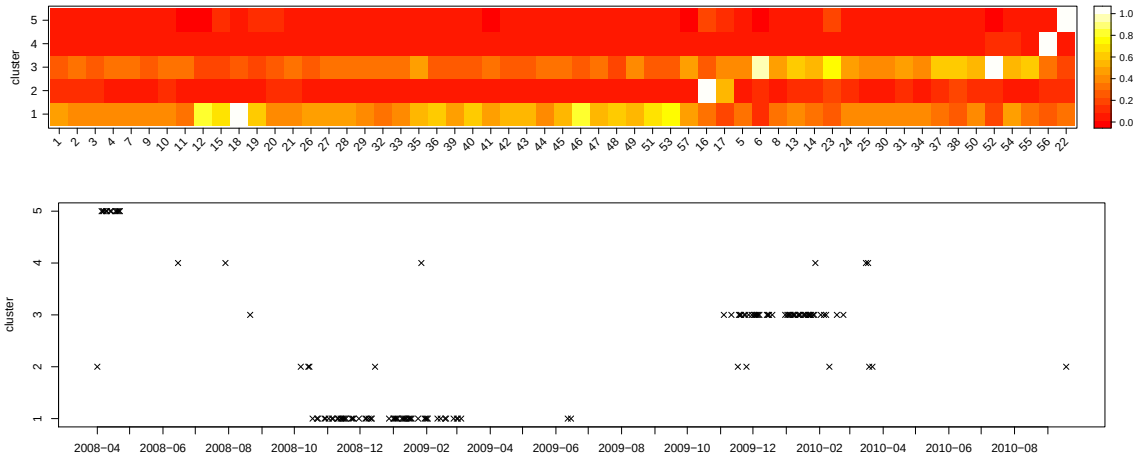


Figure 12: Clustering results of the electricity consumption using non-extremal data for $k = 5$.

From the heat maps in Figure 12 is observed that Cluster 1 focuses on young multi-ethnic living in converted flats. Cluster 2 focuses on young prosperous living in flats. Cluster 3 focuses on families and single parent living in council flats and farming communities. Cluster 4 focuses on multi-ethnic families in crowded flats and Cluster 5 focuses on low income singles. The plot in Figure 12 shows that Cluster 1 used large amount of electricity in the winter of 2008-2009 and in June 2009. Cluster 3 consumed large amount of electricity in August 2008 and the winter of 2009-2010 and Cluster 5 in April 2008. The other clusters consumed large amount scattered over the years. Similar results are obtained for the analysis of $k = 10$ in Figure 13. It is observed that clusters that appeared in the previous analysis had the same consumption pattern.

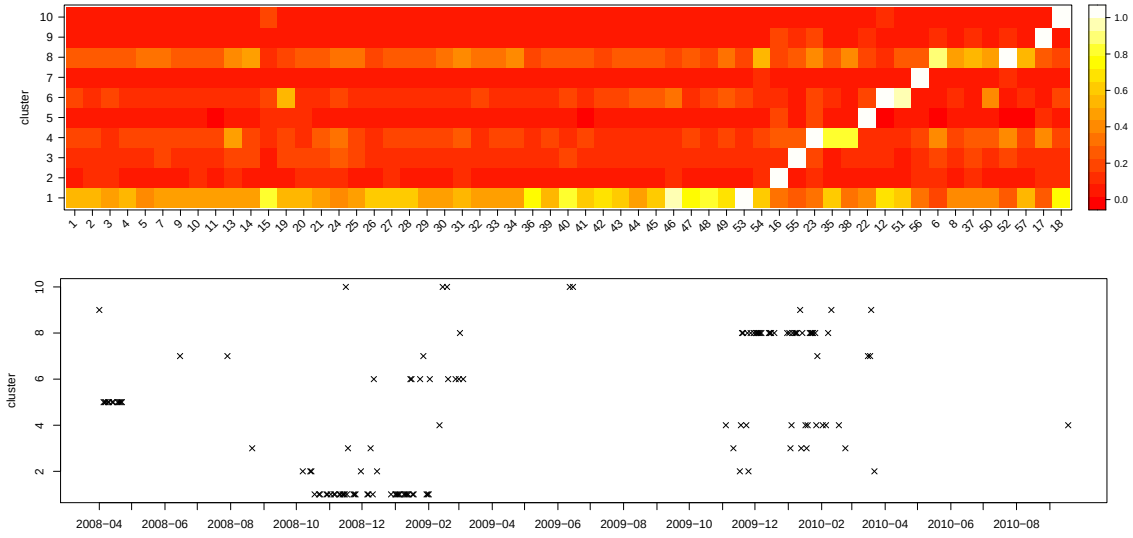


Figure 13: Clustering results of the electricity consumption using non-extremal data for $k = 10$.

6 Conclusion

In this paper, we investigated the dependence structure of extremal and non-extremal data using spherical k-means. In addition, these results are compared to find if similar patterns occur. We show using three data examples how the dependence structure of the cluster centers can be interpreted. A cluster with one large component hints at independence between all other components. A cluster with multiple large components hints at dependence.

So is observed for the extreme financial portfolio losses that the industries can be grouped in sectors: the consumer oriented, the business and IT related, the energy and materials, and industries linked to the financial sector. Meanwhile, the clusters for the non-extremal data show strong independence as almost all clusters contain one large component. It is observed that most of the extreme losses are made in 2000 in all industries due to the oil and financial crisis, the dot-com bubble, the turn of the millennium and the rise of alternative sources of energy. For the non-extremal data, losses are scattered throughout the years. However, most of the losses are concentrated around 2000.

For the extreme dietary intake data, significant clusters are formed by vitamin B2, vitamin B6 and niacin, by vitamin K and lutein, by carbs and sugar, by folate and folic acid, by vitamin B1 and iron and by fat and fatty acids which are accompanied by high intake of calories. The non-extremal data show only a few clusters: cholesterol and choline, and vitamin K and lutein. Therefore the only matching clusters between extremal and non-extremal data is the cluster vitamin K and lutein. The dependence structure of the extremal data differs from the dependence structure of the non-extremal data.

For the extreme electricity consumption data, significant clusters are formed by wealthy families with larger houses, by older households living in detached homes, by young households and by most of the communal population. Extreme electricity consumption during 2008-2010 are caused by cold spells, heavy snowfall, extreme rainfall and above average temperatures. For the non-extremal data, these types of households are not grouped together. The clusters focuses on young multi-ethnics, young prosperous, families and singles, and multi-ethnic families living in flats. It also focuses on farming communities and low income singles. The dependence structure and the consumption pattern of clusters using extremal data contrast from the non-extremal data.

From analyzing these datasets, we conclude that this procedure reveals relevant patterns and is able to visualize and classify extremal events. Furthermore, the revealed dependence structure of extremes differs from the dependence structure of non-extremes. This indicates that non-extremal behavior does not capture the behavior of extremes.

There are some limitations for the interpretations of the results. The electricity consumption dataset lacked extensive information about households. After cleaning the dataset, the electricity consumption only recorded two years of data. Therefore there are not enough observations namely winters to draw conclusions. For further research a dataset with a larger horizon and more extensive household information like household size, isolation and house sizes are useful to take into account. Furthermore, the threshold for selecting the extremal and non-extremal data and the optimal number of clusters k are crucial for the results. Therefore the robustness of the results need to be checked with these parameters. Further research could be conducted to provide insights on datasets with higher dimensions. For example dimension reduction techniques in combination with clustering in Chautru (2015) and Cooley and Thibaud (2019). An extended methodology in Fomichov and Ivanovs (2020) is given that showed that the procedure by Janßen and Wan (2020) which is applied in this paper is sub-optimal.

References

AECOM Building Engineering. (2018). Energy demand research project: Early smart meter trials, 2007-2010. https://doi.org/10.5255/UKDA-SN-7591-1

Balkema, A. A., & De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, 792–804.

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

Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic journal of statistics*, *9*(1), 383–418.

Chiapino, M., & Sabourin, A. (2016). Feature clustering for extreme events analysis, with application to extreme stream-flow data. *International Workshop on New Frontiers in Mining Complex Patterns*, 132–147.

Coles, S., Heffernan, J., & Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, *2*(4), 339–365.

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.

Engelke, S., & Ivanovs, J. (2020). Sparse structures for multivariate extremes. *Annual Review of Statistics and its Application*, *8*.

Engelke, S., & Volgushev, S. (2020). Structure learning for extremal tree models. *arXiv preprint arXiv:2012.06179*.

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical proceedings of the Cambridge philosophical society*, *24*(2), 180–190.

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

Gissibl, N., Klüppelberg, C. et al. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, *24*(4A), 2693–2720.

Goix, N., Sabourin, A., & Clémençon, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis*, *161*, 12–31.

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

Jung, S., Dryden, I. L., & Marron, J. S. (2012). Analysis of principal nested spheres. *Biometrika*, *99*(3), 551–568.

Li, D. X. (2000). On default correlation: A copula function approach. *The Journal of Fixed Income*, *9*(4), 43–54.

Pickands III, J. et al. (1975). Statistical inference using extreme order statistics. *Annals of statistics*, *3*(1), 119–131.

Salmon, F. (2012). The formula that killed wall street. *Significance*, *9*(1), 16–20.

Thibaud, E., Mutzner, R., & Davison, A. C. (2013). Threshold modeling of extreme spatial rainfall. *Water resources research*, *49*(8), 4633–4644.

Appendix

Table 4: ACORN type information from 2003.

Type	Description	% of UK population	Type	Description	% of UK population
1	Wealthy mature professionals, large houses	1.7	30	Established home-owning workers	2.6
2	Wealthy working families with mortgages	1.5	31	Home-owning asian family areas	1.1
3	Villages with wealthy commuters	2.7	32	Retired home owners	0.9
4	Well-off managers with larger houses	2.6	33	Middle-income, older couples	3.0
5	Older affluent professionals	1.8	34	Lower incomes, older people, semis	2.1
6	Farming communities	2.0	35	Elderly singles, purpose-built flats	0.7
7	Old people, detached homes	1.9	36	Older people, flats	1.9
8	Mature couples, smaller detached homes	2.0	37	Crowded asian terraces	0.5
9	Older families, prosperous suburbs	2.1	38	Low income asian families	1.1
10	Well-off working families with mortgages	2.3	39	Skilled older families, terraces	2.8
11	Well-off managers, detached houses	3.7	40	Young working families	2.1
12	Large families and houses in rural areas	0.6	41	Skilled workers, semis and terraces	3.3
13	Well-off older professionals, larger houses and converted flats	0.9	42	Home-owning families, terraces	2.8
14	Older professionals in suburban houses and apartments	1.4	43	Older people, rented terraces	1.8
15	Affluent urban professionals, flats	1.1	44	Low income larger families, semis	3.3
16	Prosperous young professionals, flats	0.9	45	Low income older people, smaller semis	3.0
17	Young educated workers, flats	0.6	46	Low income routine jobs, terraces and flats	1.4

Type	Description	% of UK population	Type	Description	% of UK population
18	Multi-ethnic young, converted flats	1.1	47	Low income families, terraced estates	2.6
19	Suburban privately renting professionals	0.9	48	Families and single parent, semis and terraces	2.1
20	Student flats and cosmopolitan sharers	0.6	49	Large families and single parents, many children	1.7
21	Singles and sharers, multi-ethnic areas	1.6	50	Single elderly people, council flats	1.8
22	Low income singles	1.2	51	Single parents and pensioners, council terraces	1.9
23	Student terraces	0.4	52	Families and single parents, council flats	0.8
24	Young couples, flats and terraces	1.0	53	Old people, many high rise flats	0.8
25	White collar singles and sharers, terraces	1.4	54	Singles and single parents, high rise estates	0.9
26	Younger white collar couples with mortgages	1.9	55	Multi-ethnic purpose-built estates	1.1
27	Middle-income, home owning areas	2.9	56	Multi-ethnic, crowded flats	1.1
28	Working families with mortgages	2.6	57	Mainly communal population	0.3
29	Mature families in suburban semis	3.3			

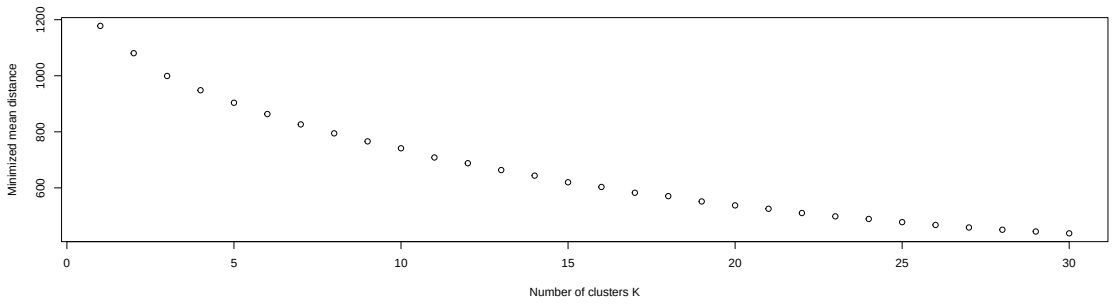


Figure 14: Elbow plot of the financial portfolio losses using non-extremal data.

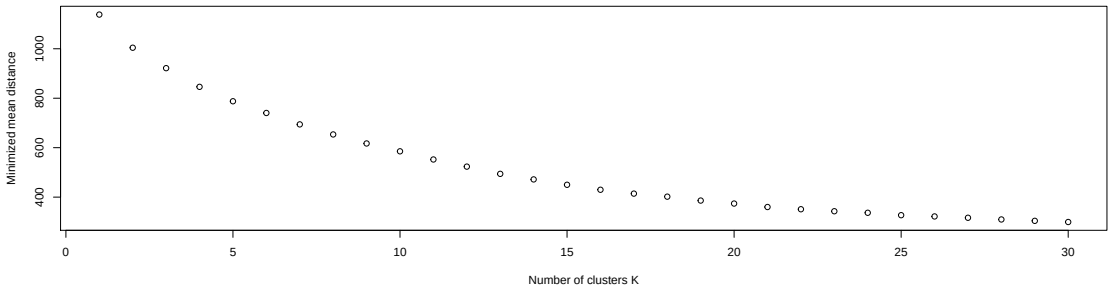


Figure 15: Elbow plot of the dietary intake using non-extremal data.

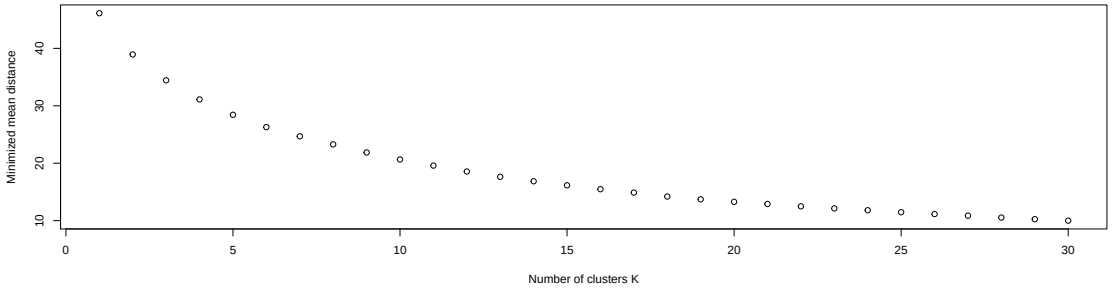


Figure 16: Elbow plot of the electricity consumption using non-extremal data.