Erasmus University Rotterdam

Bachelor Thesis

Extreme value clustering and dimensionality reduction with application in finance and beyond

Author

Assing Yang (433 991)

Supervisor

P. Wan

July, 2021

Abstract

When dealing with high dimensional sets of data, it is is often desired to reduce its complexity which assists in the interpretation of the variables. Therefore, clustering is often desired to partition large numbers of observations into several distinct groups to analyze them separately. Similarly, dimensionality reduction is also desirable to prevent data sparsity. The k-means clustering algorithm and its variant, the spherical k-means are one of the most important methods for pattern detection. While the most straightforward method to reduce dimensions is principle components analysis (PCA). In this paper, the spherical k-means algorithm will be applied to analyze the extremal observations from a data set. By making use of multivariate extreme value analysis it will show how 'extremal prototypes' can be found through clustering. An evaluation of the performance of PCA against a more sophisticated method for spherical structure, principal nested spheres(PNS) will also be given.

Contents

1 Introduction

Extreme events that have low probabilities of occurrence, but are catastrophic, whether it is environmental disaster like flood, earthquakes or financial crisis caused by underestimation of the underlying risks. Hence, a proper modelling of these events are crucial. To quantify the risks of these few very rare occurring events, extreme value theory (EVT) has been a widely-used approach as they provide a justified mathematical toolkit to estimate periods of returns of events that still have yet to be observed.

Multivariate extreme value theory (MVET) provides us with tools to model multivariate extreme dependencies. However, the amount of data is ever-increasing, although this helps tremendously when evaluating events or company decisions, the sheer amount of data can, however, be overwhelming. Therefore, reduction of dimensions is often needed to interpret the data in a meaningful way. It partially removes multi-collinearity which in turn improves the interpretation of the parameters of the model.

While dimensionality reduction gives us insight into variables, clustering provides us insight into observations. By partitioning the data set into a number of distinct groups, clustering gives great insight into the structure of the data.

MVET combined with dimensionality reduction and clustering gives us great diagnoses in financial markets for example. In financial distress analysis, it is of crucial interest to prevent heavy portfolio losses. MVET can also be very useful in health care, for analyzing rare causes for deadly diseases. In this paper we will apply MEVT on a financial portlofio dataset and a dietary dataset to find out what factors will mostly cause extreme events, and their dependencies.

2 Literature

An important aspect in EVT is the identification of limiting distributions for maxima, this idea was pioneered by Fisher and Tippett, 1928, Tippet obtained three asymptotic limiting distributions of extremes assuming independent variables, one of the distributions was identified earlier by Fréchet, 1927 and is called after him as Fréchet distribution, Fréchet, 1927 also introduced a functional equation called stability postulate and now referred to as max-stability. The work of Fisher and Tippett, 1928 will then be further developed by E. Gumbel, 1935 as he identified another form of distribution which was called after him as well and known the Gumbel distribution, Mises, 1936 and Gnedenko, 1943, resulting in the perfection of one of the most important result in EVT known as Fisher–Tippett–Gnedenko theorem in 1943. Then E. J. Gumbel, 1958 wrote the first book to exclusively evaluate the relevance of extreme values, and its application to engineering, Embrechts et al., 1997 provided an orientation of EVT towards finance and insurance.

For the multivariate extreme value theory (MEVT) however, several difficulties arose as it is not clear how to specify conditions that constitute extreme events, this problem was discussed in Morton and Bowers, 1996, Beirlant et al., 2004 provided a wider illustration of probabilistic aspects of EVT, and stated that the unknown multivariate distribution function are bounded to certain constraints Coles, 2001, and Engelke and Ivanovs, 2021 reviewed some asymptotic tools and constraints for multivariate extreme value distribution by multivariate regular variation theory introduced by Resnick, 1987. MVET has its wide applications, for example in finance Aslanertik et al., 2017. The research in MVET has been very active over the years, however most applications are still restricted to fairly modest dimensions due to a lack of clear notions of sparsity. In attempt to reduce complexity in extremal dependencies, one can apply dimensionality reduction techniques and clustering.

In clustering the idea is to find data points that have certain similarities or dissimilarities. The data points that share similar characteristics are then stored in the same cluster. This way we can simplify the data by dividing the total number of observations into lower subsets. According to Mirkin, 1998 the most often used clustering algorithm is the k-means method, first proposed by MacQueen, 1967. K-means starts with a set of some chosen cluster centers, and tries to form clusters around these centers by minimizing the sum of all euclidean distances between data points and the chosen cluster centers while keeping the centers updated regularly in each iteration. However, k-means is often sub-optimal, as the starting points that are chose as centers have huge impact on the results, where it is possible for the algorithm to end in a local minimum.

In Dhillon, 2001, a k-means clustering method for unstructured text data was presented, in which they had created a *vector space model* for text data and extracted unique content-bearing words from the set of documents and classified those words as 'variables' or 'features'. Each document was then represented as a vector of word frequencies in the vector space. Further, in the paper a cosine similarity was used instead of the original Euclidean distance, and this adapted version of k-means was referred to as spherical k-means.

In order to to build appropriate models and acquire accurate results, scientist are attracted to add as many features as possible to the data. However, having too many variables in the data causes 'the curse of dimensionality'. The concept of 'curse of dimensionality was first by Bellman, 1957, when examining problems regarding dynamic programming. Formally, the curse of dimensionality refers to the occurrence of sparsity in data caused by the too rapid increase of space volume when adding new explanatory variables to the data. This sparsity forms a barrier on the way of obtaining reliable statistical significance .

To avoid the curse of dimensionality, and mitigate its effects, it is crucial that dimensionality reduction is applied. The techniques often fall into two categories: feature selection and feature

extraction. Methods regarding feature selection are discussed in Isabelle Guyon, 2003 extensively. This paper will mainly focus on feature extraction. With feature extraction, variables existing in the high dimensional space, are combined into components in a lower dimensional space. This is done in such a way that maximum variance is preserved.
The most classical linear convex DR method is Principal Component Analysis (PCA), in which high dimensional correlated data is transformed to a lower dimensional set of uncorrelated components, this method was first proposed by Pearson, 1901.

3 Data

The methods will be applied on two different kinds of data. Firstly, a financial loss portfolio is used, this data was discussed in Cooley and Thibaud, 2019 where a method related to Principal Component Analysis (PCA) was used to examine extreme losses dependencies. Secondly, a dietary intakes data will be explored. In this the financial portfolio losses data, the 'value-averaged' daily returns of 30 industry portfolios are assembled, the data has a total of 16694 observations and spans from 1950 to 2015.This data was obtained through Kenneth French Data Library.

The dietary data was obtained from ://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DR1TOT_I.XPT, and contains dietary interviews from 2015-2016 NHANES report. The interview recorded all the consumption of food and beverage from participants during the 24 hours prior to the interview. As result the dataset contains information of the nutrients from the observations. Our interest lies in exploring the dependencies of 38 chosen nutrients that are taken in on a high level, as negative health effects might occur by high doses of some components.

4 Methodology

4.1 Extreme value theory

To capture extreme deviations from the median of probability distributions, it is essential to derive probability distributions for the 'extreme tail losses', this can arguably be done through central limit theorem, in which the convergence behaviour of random variables are examined to find their limiting distribution. The same principle can be followed in order to find distributions for the maximum of random variables.

4.1.1 Univariate extreme value theory

Let $X_1, \ldots, X_n$ be i.i.d. random variables, with the corresponding probability distribution function (PDF) $f_X(x)$ and cumulative distribution function (CDF) $F_X(x)$. Let $M_n = max(X_1, ..., X_n)$ be the maximum of the random variables. Theoretically, since the variables are i.i.d. the distribution of this maximum can be derived as follows when having an extreme event z:

$$F_{M_n}(z) = P(M_n \leq z) = P(X_1 \leq z) \cdots P(X_n \leq z) = (F_X(z))^n$$

Hence the probability $p(z)$ that an extreme event occurs is equal to $1 - (F_X(z))^n$. However, the theoretical distribution $F_X(x)$ is often unknown, thus only the asymptotic distribution can be derived. Fisher & Tippett (1928) had shown in their results that when the maximum of a sample of i.i.d. random variables is linearly rescaled, it can only converge in distribution to certain distribution G that is formed by either Gumble distribution, Fréchet distribution or Weibull distribution. That is, for any existing sequences of real numbers $a_n > 0$ and $b_n \in \mathbb{R}$, F_X is in the max-domain of attraction of the extreme value distribution G ($F \in \text{MDA}(G)$):

$$\lim_{n\to\infty} (F_X(a_n z + b_n))^n = \lim_{n\to\infty} P\left(M_n \leq a_n z + b_n\right) = \lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq z\right) = G_Z(z)$$

We can also say that $(M_n - b_n)/a_n \overset{d}{\to} Z$ as $n \to \infty$ where $G_Z(z)$ is a non-degenerating limiting distribution, by merging the aforementioned three forms of G, the generalized extreme value distribution (GEV) can be found:

$$G(z; \mu, \sigma, \xi) = \begin{cases} \exp\Big\{ - \big[1 + \xi(\frac{z-\mu}{\sigma})\big]^{-\frac{1}{\xi}} \Big\} & , \quad \xi \neq 0 \quad and \quad 1 + \xi(\frac{z-\mu}{\sigma}) > 0 \\ \exp\Big\{ - \exp\big[-(\frac{z-\mu}{\sigma})\big] \Big\} & , \quad \xi = 0 \end{cases} \tag{1}$$

The corresponding PDF $g(z)$ can then be obtained by differentiating $G(z; \mu, \sigma, \xi)$ and then be found as:

$$\frac{1}{\sigma} h(z)^{\xi+1} e^{-h(z)} \quad where \quad h(z) = \begin{cases} \big[1 + \xi(\frac{z-\mu}{\sigma})\big]^{-\frac{1}{\xi}} & if \quad \xi = 0 \\ \exp\big[-(\frac{z-\mu}{\sigma})\big] & if \quad \xi = 0 \end{cases} \tag{2}$$

The parameters ξ, μ and σ are shape, location and scale parameters respectively, the parameter ξ determines the weight of the uppertail of the density. When ($\xi > 0$) the GEV is heavy-tailed and follows the Fréchet distribution and is also know is type 2, if ($\xi < 0$) the distribution is light-tailed and has a Weibull form and is known as type 3, in case ($\xi = 0$), the distribution has an exponential tail and follows Gumbel law and is known as type 1.

4.1.2 Multivariate extreme value theory

In order to study the extremal behaviors of random vectors, the most natural assumption is to study the limiting distributions of the componentwise maxima. This is done by finding the marginal distributions and extremal dependence. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random vectors, and $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})$ for $1 \leq i \leq n$ *and* $i, d \in \mathbb{N}$. Then we can construct a vector $\mathbf{M}_n = (M_{n,1} = \max_{i=1,\ldots,n} X_{i,1}\ , \ldots,\ M_{n,d} = \max_{i=1\ ,\ldots,\ n} X_{i,d})$ which consists of componentwise maxima, by following the theory for univariate random variables, there exists sequences of constants $\{a_{n,j}\} > 0$ and $\{b_{n,j}\} \in \mathbb{R}$ with $1 \leq j \leq d$ such that when properly linear normalized, all maxima converge to a jointly non-degenerate limiting distribution:

$$\lim_{n\to\infty} P\left(\frac{M_{n,1} - b_{n,1}}{a_{n,1}} \leq z_1, \ldots, \frac{M_{n,d} - b_{n,d}}{a_{n,d}} \leq z_d\right) = G(z_1, \ldots, z_d) \tag{3}$$

With the convergence $(M_{n,j} - b_{n,j})/a_{n,j} \xrightarrow{d} Z_j$ as $n \to \infty$, and it's corresponding CDF and PDF for the marginal distribution $G_j(z)$ can be found in equation (1) and equation (2) with parameters (u_j, σ_j, ξ_j) and $1 \leq j \leq d$.

Now, the focus is on characterizing extremal dependencies between random variables. Let F_j be the marginal distribution function of X_j with $1 \leq j \leq d$, without loss of generality we assume that X_j is theoretically standard Pareto distributed such that $F_j(x) = 1 - 1/x$. Such that X_j is transformed to $1/\{1 - F_j(X_j)\}$, this will scale all the components and give meaning to the understanding of 'large' values. By normalization the $M_{n,j}$ while choosing for $a_{n,j} = n$ and $b_{n,j} = 0$ we can derive the following:

$$\begin{aligned} \lim_{n\to\infty} P\left(\frac{M_{n,j}}{n} \leq z_j\right) &= \lim_{n\to\infty} P\left(M_{n,j} \leq nz_j\right) = \lim_{n\to\infty} \left(F_j(nz_j)\right)^n \\ &= \lim_{n\to\infty} \left(1 - \frac{z_j^{-1}}{n}\right)^n = \exp\{-z_j^{-1}\} \end{aligned}$$

Thus,

$$G_j(z_j) = \exp\{-z_j^{-1}\}, \qquad z_j \in [0, \infty) \tag{4}$$

We can see that the marginal distribution for Z_j is standard Fréchet with $\xi_j = \sigma_j = 1$ and $\mu_j = 0$. for $1 \leq j \leq d$. Then, the so called multivariate extreme-value distribution $G(z_1, \ldots, z_d)$ is max-stable and can be expressed as :

$$G(z_1, \ldots, z_d) = \exp\{-V(z_1, \ldots, z_d)\},\ \textit{with}\ \mathbf{z} = (z_1, \ldots, z_d) \in \mathbb{R}_+^d \tag{5}$$

where the function $V(z_1, \ldots, z_d)$ is called the underling exponent measure, this exponent measure can be any positive function that is homogeneous with degree of -1 (i.e.. that $V(rz_1, \ldots, rz_d) = r^{-1}V(z_1, \ldots z_d)\ \forall c > 0$) and suffices the marginal constraints $V(z, \infty, \ldots, \infty) = \frac{1}{z}$ for all permutations of indices. For the convergence in equation (3) to hold, it is said that the vector $\mathbf{X}$ is required to be multivariate regular varying. $\mathbf{X}$ is multivariate varying if there exist function $a(\cdot)$ with $\lim_{t\to\infty} a(t) = \infty$ and a measure V on all Borel sets $A \subset \mathbb{E} = [0, \infty)^d \setminus \{\mathbf{0}\}$ such that

$$\lim_{t\to\infty} tP\left(\frac{\mathbf{X}}{a(t)} \in A\right) = V(A) \tag{6}$$

Since the exponent measure is homogeneous, it is often more convenient to transform the vectors to polar coordinate, for this we will use some arbitrary but fixed norm $\|\cdot\|$ and apply polar coordinate transformation. First we define a positive unit sphere $\mathbb{S}_+^{d-1} = \{\mathbf{x} \in \mathbb{E} : \|\mathbf{x}\| = 1\}$. Then we define the transformation $T : \mathbb{R}^d \setminus \{\mathbf{0}\} \to (0, \infty) \times \mathbb{S}_+^{d-1}$ by:

$$T(\mathbf{x}) = \left(\|\mathbf{x}\| \ , \ \frac{\mathbf{x}}{\|\mathbf{x}\|} \right) =: (r, \boldsymbol{\theta}) \tag{7}$$

Where $\boldsymbol{\theta}$ is the corresponding angle. Then in equivalence to equation (6), (define $a(t) = t$ for convenience) there exists a probability measure $S(\cdot)$ on space $\mathbb{S}_+^{d-1}$ such that for $T(\mathbf{X}) = \left(\|\mathbf{X}\| , \frac{\mathbf{X}}{\|\mathbf{X}\|} \right)$ and all Borel sets $B \subset \mathbb{S}_+^{d-1}$

$$\lim_{t\to\infty} P\left(\frac{\mathbf{X}}{\|\mathbf{X}\|} \in B \mid \|\mathbf{X}\| > t \right) = P(\boldsymbol{\Theta} \in B) = S(B) \tag{8}$$

Where $(\mathbf{X} / \|\mathbf{X}\|)$ given a norm that exceeds some high thresholds weakly converges to $\boldsymbol{\Theta}$ and follows a spectral distribution for which its cumulative mass is measured by $S(\cdot)$.

Since our exponent measure V is homogene, it will decompose $\mathbf{z} = (z_1, \ldots, z_d)$ into an independent radial part and angular part, the expression for the depending structures of the maxima can then be expressed as

$$V\{\mathbf{z} \in \mathbb{E} : \|\mathbf{z}\| \geq y \ , \ \mathbf{z}/\|\mathbf{z}\| \in B\} = cy^{-1} S(B) \quad \forall y \in \mathbb{R}_+ \tag{9}$$

with $S(B)$ being the angular part of the expression and describes the dependence structure of maxima. $c = V(\mathbf{z} \in \mathbb{E} : \|\mathbf{z}\| \geq 1)$ is fixed and constant. With $S(B)$ is the angular part of the expression. By decomposing into pseudo-polar coordinates a spectral representation for the exponent measure can also be given:

$$V(z_1, \ldots, z_d) = D \int_{\mathbb{S}_+^{d-1}} \bigvee_{j=1}^{d} \left(x_j z_j^{-1} \right) \mathrm{d}S(x_1, \ldots, x_d) (10)$$

Where $\bigvee$ denotes for maximum, and the spectral distribution S must satisfy the mean constraint:

$$\int_{\mathbb{S}_+^{d-1}} x_j \, \mathrm{d}S(x_1, \ldots, x_d) = D^{-1} \quad j = 1, \ldots, D \tag{11}$$

The spectral measure $S(\cdot)$ gives us information about the angle or direction of an extreme observation, a small set under $S(\cdot)$ but with a high probability can be considered as a direction for an extreme event, in Janßen and Wan, 2020, they identified these extremal patterns in a non-parametric way, by identifying these small sets without assuming a specific model.

4.2 Max-linear model

The simplest classical non-parametric method for modeling extremal dependency under spectral measure is the so called max-linear model. A max-linear model consists of k different non-negative coefficient vectors $\{\mathbf{a}_1, \cdots, \mathbf{a}_k\}$ with $\mathbf{a} \in [0, \infty)^d$. Let $Z_i,\ i = 1, \cdots, k$ be random standard Fréchet distributed variables, then a entry X_i of a random d-dimensional vector $\mathbf{X}$ can be expressed as :

$$X_i = \max\left(a_1^i Z_1, \cdots, a_k^i Z_k\right) \qquad for\ all\ i = 1, \cdots, d \tag{12}$$

Further we assume that:

$$\sum_{i=1}^{k} a_i^j = 1 \qquad for\ all\ j = 1, \cdots, d \tag{13}$$

We can clearly see from equation (12) that by having large Z_i, the value for X_i will become extreme. Hence we can say that the possible directions of a extreme observations can be determined by the coefficient vectors $\{\mathbf{a}_1, \cdots, \mathbf{a}_k\}$. Equivalently, the spectral measure is concentrated around k points, and the angle $\mathbf{\Theta}$ can only take values $\mathbf{a}_i / \|\mathbf{a}_i\|$ with the corresponding probability $\|\mathbf{a}_i\| / (\sum_{j=1}^{k} \|\mathbf{a}\|_j)$, for $i = 1, \cdots, d$. An estimation for the max-linear model can be done through clustering which is explained in the next section.

4.3 Estimation of Max-linear model through Clustering

Clustering based on centroids are aiming to find a set of k points $(c_1, \ldots, c_k) \in \mathbb{R}^d$ such that the following objective function is minimized:

$$\mathbf{min}_{j=1,\cdots,k}\ d(\Theta, c_j) \tag{14}$$

For any random object $\Theta \in \mathbb{R}^d$ and a dissimilarity fucntion d. In our case the random object Θ is an extremal angle that is appearing in the decomposition of the exponent measure V in equation (9) with the spectral distribution S. By decomposing the spectral measure into k different clusters, the angular distribution will concentrate around k number of points in $\mathbb{S}_+^{d-1}$. Hence, this can also be seen as an estimation of the coefficients in a max-linear model. In Janßen and Wan, 2020 the spherical k-means clustering method was proposed for clustering the angle $\mathbf{\Theta}$.

4.3.1 Spherical k-means clustering

The original k-means algorithm was introduced by MacQueen, 1967 and works as follows: consider a dataset with n observations is given, where each observation is a multidimensional vector. The aim is to partition the data into K clusters, with the obvious restriction that $K \leq n$. Now for each cluster C_j such that $j = 1, ..., K$, let $\{\mathbf{m_j}\}_{j=1}^{k}$, $\mathbf{m_j} \in \mathbb{R}^d\ j \in \mathbf{N}$ be a randomly selected cluster centroids. For each observation, the distance to each of the respective cluster centroids is minimized. And subsequently the observation is assigned to the cluster that is "closest".

More formally, we want to define and minimize the distance of each object to the nearest centroid by the means of a distance/dissimilarity function $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$. Since we are interested in finding k centroids $\{\mathbf{m_j}\}_{j=1}^{k}$ in a unit sphere $\mathbb{S}_{+}^{d-1}$ such that the expected dissimilarity of X and the closest centroid is minimized, it is more natural to dissimilar two random vectors by their angles, instead of using the classic Eulicdean distance $\|\mathbf{x} - \mathbf{y}\|_2$. This is a variant of k-means and is known as spherical k-means procedure introduced by Dhillon, 2001. The dissimilarity function $d(\cdot)$ is defined as

$$d(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y}) = 1 - \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} \tag{15}$$

The objective function of the standard spherical k-means can then be formulated as

$$\mathbf{min} \sum_{i,j} \mu_{ij} (1 - \cos(\mathbf{x}_i, \mathbf{m}_j)) \tag{16}$$

Where

$$\mu_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ is assigned to cluster} j \\ 0 & \text{otherwise} \end{cases}$$

And u_{ij} is element of the binary membership matrix $\mathbf{U} \in \mathbb{R}^{n \times k}$, and if we hold strict to the unit sphere $\mathbb{S}_{+}^{d-1}$ then the euclidean norms are equal to 1 and $1 - \cos(\mathbf{x}_i, \mathbf{m}_j) = 1 - \langle \mathbf{x}_i, \mathbf{m}_j \rangle$.

4.3.2 Elbow method for determining optimal k

The k-means algorithm aims to partition the data into k distinct clusters by the means of their angles. Since$k$ is fixed and predetermined, a suitable value needs to be found before the algorithm can be used. This can be done using the *within-cluster sum of squares* (WSS). In this paper, instead of the WSS that is most applicable to Euclidean k-means, a more general formulation of it is given as the averaged distance from any observation to the closest element of the cluster centers. That is, assuming that S is a on $\mathbb{B}(\mathbb{S}^{d-1})$, and that we have a set $A = \{\mathbf{a}_1, \cdots, \mathbf{a}_k\}$, then the average distance from any observation x to the closest element of A is :

$$W(A, S) = \int_{\mathbb{S}^{d-1}} \min_{\mathbf{a} \in A} d(\mathbf{x}, \mathbf{a}) S(d\mathbf{x}) \in [0, \infty) \tag{17}$$

For a given k, the minimization of $W(A, P)$ with respect to A can be see as the process of finding cluster centers, where the set of k cluster centers is defined as A_k. If we use a measure that places mass $1/n$ on each observation of a sample, and denote it as S_n. and an optimal set of A_k^n for this measure is derived. Then $W(A_n^k, S_n)$ is the minimized average distance from any observation it its cluster center for k number of clusters. The number of clusters is considered optimal when the marginal decrease in $W(A_n^k, S_n)$ starts to increase.

4.4 Dimensionality reduction

When modeling data with high dimensions, it becomes more difficult to cluster data properly. This phenomenon is known as the curse of dimensionality. Hence dimensionality reduction (DR) techniques are crucial. The most well-known DR method was introduced by by Pearson, 1901 and is known as principal component analysis (PCA), which makes uses of finding principal geodesics on a Euclidean plane, that is the shortest path between two points expressed in a straight line.
However, since we are working with angles in a unit sphere $\mathbb{S}^{d-1}$, it is more appropriate to make use of great cycles instead of straight lines as geodesic distance. In JUNG et al., 2012, a general framework for a decomposition of a high-dimensional sphere was proposed, in which they used the analysis of principal nested spheres (PNS) that was an extension on the PCA. The PNS provides an effective analysis of the main modes of variation of the data and its effectiveness will be compared with normal PCA to see which approach is more suitable in an extreme setting.

4.4.1 Principal component analysis

The fundamental idea of PCA is to reduce dimension of a random vector $\mathbf{X} \in \mathbb{R}^d$ by obtaining vectors, referred to as components, that capture the largest proportion of variance within a data set in as few components as possible. The acquired principal components represent a linear combination of the original variables and are orthogonal to each other.

$$PC_1 = \mathbf{a}_1\mathbf{X}_1 = \alpha_{11}x_1 + \alpha_{21}x_2 + \alpha_{31}x_3 + ... + \alpha_{d1}x_d$$

$$\vdots$$

$$PC_d = \mathbf{a}_d\mathbf{X}_d = \alpha_{1k}x_1 + \alpha_{2k}x_2 + \alpha_{3k}x_3 + ... + \alpha_{dd}x_d$$

where $\mathbf{a}^\intercal\mathbf{X}$ is the approximation of $\mathbf{X}$ that can be projected onto a p dimensional space, with $p < d$. To obtain the principal components themselves eigendecomposition of the covariance matrix $\Sigma = \mathbb{E}(\mathbf{X}\mathbf{X}^\intercal)$ is the most widely used technique for obtaining the eigenvalues and eigenvectors where $\lambda_1 = \max\{\lambda_1, \lambda_2, \cdots, \lambda_k\}$, and corresponding eigenvector $v_1, \cdots, v_d$, gives the order of the principal components. By summing up the variances described by each component, which correspond with their respective eigenvalues, the cumulative explained variance is obtained. The proportion explained through a single principal component can thus be obtained by dividing its eigenvalue by the total explained variance. One way to reduce the dimension from the data is by selecting only the first ℓ principal components which explain a certain proportion of the total explained variance.

4.4.2 Principal nested spheres

In Chautru, 2015, principal nested spheres (PNS) was suggested for dimensionality reduction under a spherical condition. PNS is an iterative algorithm that projects the data on smaller subspheres with a lower dimension, these subspheres are identified with $\mathbb{S}^{d-2}, \mathbb{S}^{d-3}, \cdots, \mathbb{S}^1$. More formally, we can say that for a unit sphere $\mathbb{S}^d$, a geodesic for two points is a great cycle joining the two points and its distance function is the geodesic distance function $d^{\ell}_{PNS}(\cdot,\cdot)$ that is defined as the length of the shortest great cycle segment. The geodesic distance between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{S}^d$ can then be expressed as $d_{PNS}(\mathbf{x}, \mathbf{y}) = \arccos(\mathbf{x}^\intercal \mathbf{y})$, the shortest great cycle path is non-unique if and only if $\mathbf{x}^\intercal y = -1$. By definition, a subsphere $A^{\ell-1}$ in $\mathbb{S}^\ell$ is explained by an axis $\mathbf{v} \in \mathbb{S}^\ell$ and radius $r \in (0, \pi/2]$ and can be expressed as follows:

$$A^{\ell-1}(\mathbf{v}, r) = \{\mathbf{x} \in S^\ell : d^{\ell}_{PNS}(\mathbf{v}, \mathbf{x}) = r\} \tag{19}$$

We can see that $A^{\ell-1} \subset \mathbb{S}^\ell \subset \mathbb{R}^{\ell+1}$ and $A^{\ell-1} \subset \{\mathbf{x} \in \mathbb{R}^{\ell+1} : \mathbf{v}^\intercal \mathbf{x} - \cos(r) = 0\}$, hence we can say that the subsphere $A^{\ell-1}$ is an intersection between between a higher dimensional unit sphere and a hyperplane, from this we can conclude that $A^{\ell-1}$ is a slicing of $\mathbb{S}^\ell$ with the hyperplane.
For any $\mathbf{v} \in \mathbb{R}^\ell$, there is a $(\ell+1) \times (\ell+1)$ rotation matrix $R(\mathbf{v})$ that moves $\mathbf{v}$ to the north pole, and a $\ell \times (\ell+1)$ matrix $R^-(\mathbf{v})$ consisting the first ℓ rows of the rotation matrix $R(\mathbf{v})$. To identify the subsphere $A^{\ell-1}$ as a unit sphere $\mathbb{S}^{\ell-1}$, a transformation $f : A^{\ell-1} \to \mathbb{S}^{\ell-1}$ can be applied. This transformation is are defined by $\mathbf{v} \in \mathbb{S}^{\ell-1}$ and radius $r \in (0, \pi/2]$ as:

$$f_{\ell-1}(\mathbf{x}) = \frac{1}{\sin(r)} \mathbf{R}^-(\mathbf{v})\mathbf{x} \qquad for\ all\ \mathbf{x} \in \mathbb{S}^{\ell-1} \tag{20}$$

In order to find the best fitting subsphere, we have to minimize the residual ξ from the subsphere $A^{\ell-1}$. The residuals can be defined as the signed length of the minimal geodesic that joins from $\mathbf{x}$ to $A^{\ell-1}$. The subsphere is best fitting if the sum of squared residuals (SSR) is minimized from all the vectors $\mathbf{x}$ to the subsphere $A^{\ell-1}$. Since the radius r from a vector $\mathbf{x}$ to the the center subsphere is the distance $d_{PNS}(\mathbf{x}, \mathbf{y})$ added with some constant ξ, The residual can be expressed as follows:

$$\xi = d_{PNS}(\mathbf{x}, \mathbf{y}) - r \tag{21}$$

And the objective function becomes:

$$\min \sum_{i=1}^{n} \xi_i^2 = \min \sum_{i=1}^{n} (d_{PNS}(\mathbf{x}_i, \mathbf{y}) - r)^2 \tag{22}$$

By extending the expression in equation 20 with other objective functions, the projection $P(\cdot)$ of any point $\mathbf{x}$ on the subsphere $A^{\ell-1}$ can be denoted by:

$$P(\mathbf{x}; A^{\ell-1}(\mathbf{v}, r)) = \frac{\sin(r)\mathbf{x} + \sin(d_{PNS}(\mathbf{x}, \mathbf{y}) - r)\mathbf{v}}{\sin(d_{PNS}(\mathbf{x}, \mathbf{y})} \tag{23}$$

Here, we denote $\tilde{\mathbf{x}} = P(\mathbf{x}; A^{\ell-1}(\mathbf{v}, r))$ for the projected $\mathbf{x}$.
The objective of PNS is to retain the largest possible total variance, in the least amount of nested spheres. We keep this in mind when deciding on the number of dimensions. By summing up the relative variances described by each nested sphere, explained variance is obtained. The relative variance $V(\ell)$ for $\ell = 1, \cdots, d-1$ for each unit sphere can be calculated as :

$$V(\ell) = \left[\sum_{i=1}^{n} \left(\xi_i^{\ell}\right)^2\right] \Big/ \left[\sum_{j=1}^{d-1}\sum_{i=1}^{n} \left(\xi_i^{j}\right)^2\right] \tag{24}$$

Where ξ_i^j is the scaled residual according to JUNG et al., 2012 obtained from the projection of observation $\mathbf{x}_i$ from sphere $\mathbb{S}^j$ onto subsphere $\mathbb{S}^j$. Combining the scaled residuals and we get a $n \times d$ matrix:

$$\tilde{\mathbf{X}}_{PNS} = \left[\mathbf{\Xi}^0 \cdots \mathbf{\Xi}^{d-1}\right] \qquad where\ \mathbf{\Xi}^{\ell} = [\xi_1^{\ell} \cdots \xi_n^{\ell}]\ for\ \ell = 0, \cdots, d-1$$

The residuals may be regarded as the PCA scores. The relative variance can take values in $(0, 1]$. Similar to PCA, the number of dimensions can be reduced by choosing a certain threshold for the total explained variance.

5 Results

In this section, findings that are obtained by using the proposed methods on two different sets of data will be presented. Firstly, the data will be transformed with the help of the empirical distribution. And only a proportion with the largest norms will be chosen. Then the chosen observations will be projected onto the unit sphere. Lastly, the spherical k-means procedure will be applied to the projected sample. Furthermore, this will be extend by using a dimensionality reduction on the sample, with a comparison between PCA and PNS. The clustering quality between the original datasets and the reduced datasets will also be analysed.

For both datasets, there were many components with values close to 0 in each estimated cluster center $\mathbf{a}_i$, $, i = 1, \cdots, k$. This could point at asymptotic independence between the components. In extreme value theory, a random variable X is said to be asymptotic independent from random variable Y if :

$$\lim_{u \uparrow 1} P(X > F_X^{-1}(u) | (Y > F_Y^{-1}(u)) = 0 \tag{25}$$

In this paper by looking at the estimated clusters and observed differences in cluster components, we are hoping to find clues about asymptotic dependence or independence.

5.1 Financial portfolio losses

For this dataset, we explore the dependencies in extremal losses. The same dataset was analyzed in Cooley and Thibaud, 2019, in which they used a method related to PCA to decompose dependencies for high dimensional extremes. In this paper another different dimensionality reduction technique will be used besides normal PCA. With a clustering approach to get more insights about the dependencies.

All the returns from the data were multiplied by -1, so that the extremal losses are found. After that, the data was transformed into Fréchet marginals by applying their empirical cumulative distribution, from the transformed dataset, only observations with the largest 5% of the Euclidean norms were kept, giving us a total of 835 observations.

First, a set of suitable values k for the optimal number of clusters must be determined. This is often done by creating an 'elbow plot'. In the elbow plot, the minimized distances $W(A_k^n, S_n)$ is plotted against k. The values $W(A_k^n, S_n)$ are strictly decreasing for a larger k, the goal is to find a a k for which the decreases become insignificant for larger values, the sudden decrease in descending speed of $W(A_k^n, S_n)$ will make the curve look like an 'elbow', therefrom the name 'elbow plot'. The plot for the financial data can be found in figure 1 below:

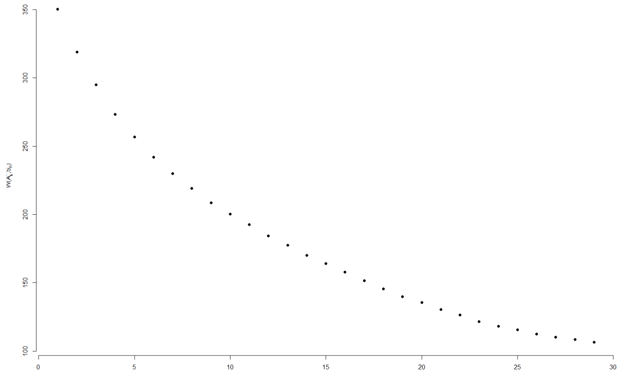


Figure 1: The value of the minimized mean distance $W(A_n^k, S_n)$ for different values of k in the financial portfolio loss data.

From the elbow plot in figure 1 no concrete values for the number of clusters could be found. As the average distance is decreasing steadily and continually over k. So we will just compare the clustering performances for $k = 5$ and $k = 10$.

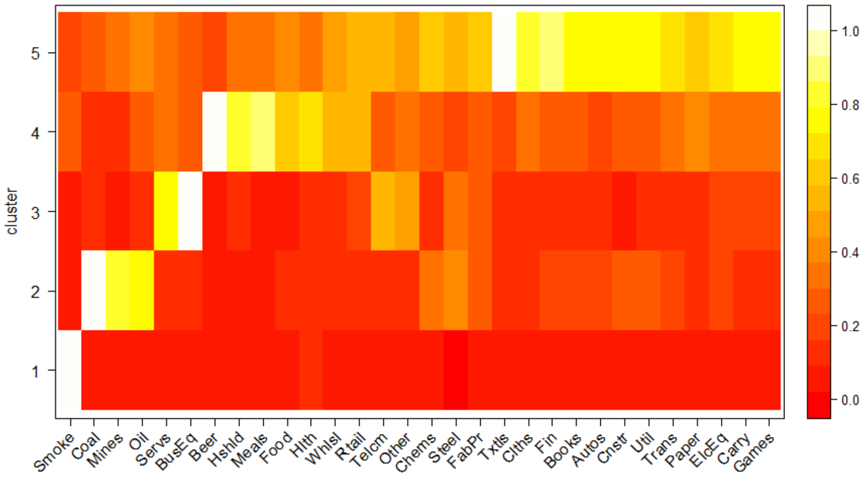


Figure 2: The k-means clustering result on the financial portfolio loss data for $k = 5$. Each row corresponds with to one of the ten estimated cluster centers, where values have been normalized as proportion of the largest variable. Which means that having a lighter color means that the corresponding component is relatively larger to all other components in the same center.

In the heat maps, the cluster centers are normalized by dividing all variables with the maximum value such that the maximum component is scaled to 1. This will provide a relative comparison between the components. In figure 2 we see the illustration of the cluster centers for $k = 5$, the clusters seem to be clearly structured and vastly different. For example, cluster 1 indicates the asymptotic independence of tobacco industry to all the other sectors, as the other components only share $0\% - 15\%$ proportionally to variable 'Smoke'. The same can be said for other cluster centers, as we see that cluster 2 is mainly driven by the coal, mineral and oils and thus energy sector. Cluster 3 focuses on Business and IT. Cluster 4 consists of consuming services and cluster 5 comprises the rest.

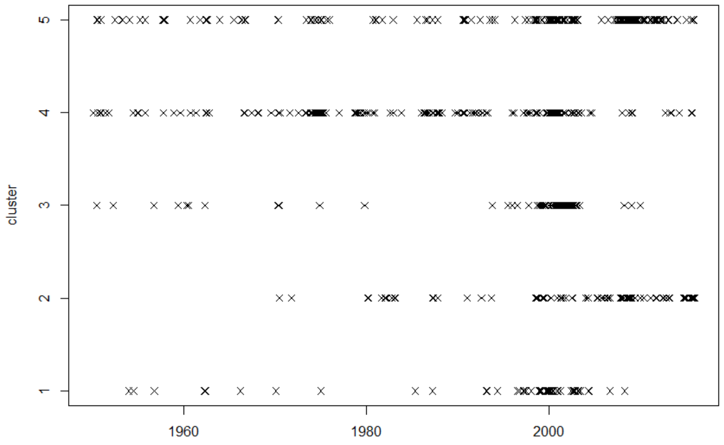


Figure 3: The k-means clustering result on the financial portfolio loss data for $k = 10$.

The density of each cluster over time.

By taking a look at figure 3, some very insightful information can be found about the behaviour from our variables. As we can see, most observations from cluster 1 are concentrated around the year 2000, in that year some severe legal actions were taken against the tobacco industry, among which a smoking ban for anyone that was born after 2000. This explains why smoking was by far the most significant in cluster 1. For cluster 2, all we know is that around 2002, the government in U.S. had released some new regulations concerning the coal mining category in order to fight against pollution. Cluster 3 indicates an internet hype started from the late nineties. The consuming goods in cluster 4 were heavily affected by the Great Recession around 2007. Cluster 5 is mainly a combination of internet hype and financial crisis.

Figure 4 shows the cluster structures for $k = 10$. The structure is very similar to clustering with $k = 5$, as most cluster centers still only have one or two relative large components, again giving a solid ground for assuming a overall asymptotic independence. The sectors which are jointly sensitive for events that cause extremal losses and move together can still easily be identified, such as the consumer service sector (cluster 5: food, Retail, consumer goods and Health), the energy sector (cluster 7: coal, mines and oil), Business and It sector (cluster 8) and the manufacturing sector (cluster 10).

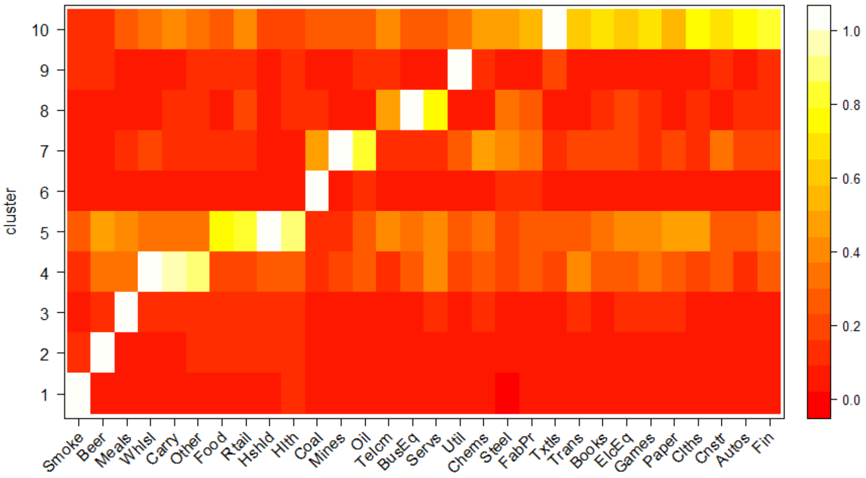


Figure 4: The k-means clustering result on the financial portfolio loss data for k = 10. Each row corresponds with to one of the ten estimated cluster centers, where values have been normalized as proportion of the largest variable. Which means that having a lighter color means that the corresponding component is relatively larger to all other components in the same center.

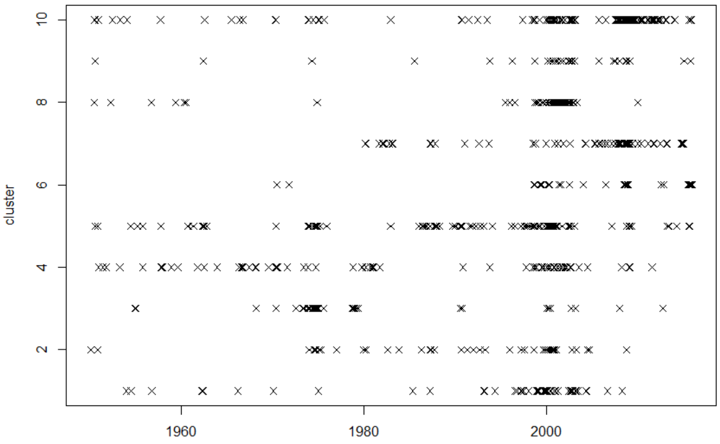


Figure 5: The k-means clustering result on the financial portfolio loss data for k = 10.

The density of each cluster over time.

The cluster over time plot from figure 5 for 10 clusters resembles a lot of figure 3, identifying important historical events in the economy.

5.1.1 PCA and PNS

Next we proceed with the PCA for the data. With the use of the *prcomp* function in *R* the principal components can then be found in figure 6 in which we present the explained variance by all the 30 components:

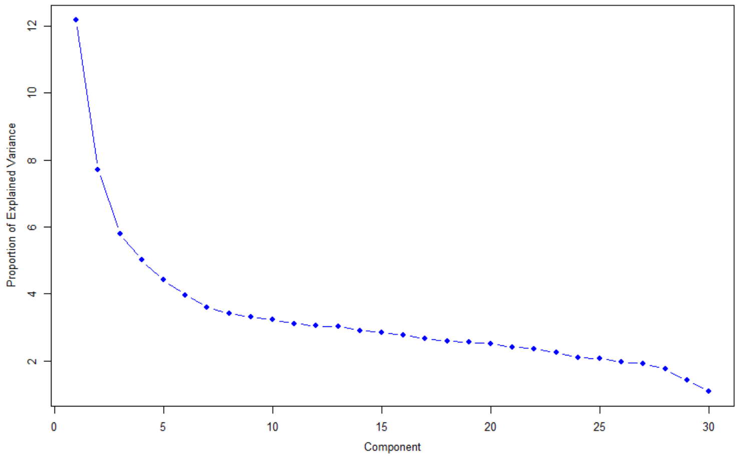


Figure 6: First 30 principal components

As expected, the variance of each principal component decreases as we go further down the horizontal axis. Noteworthy is that the first component only captures 12.18% of the total explained variance. The proportional variance thereafter sharply drops down to 7.71% We set the threshold of total explained variance to be 90% and found out that the first 25 principal components sum up to a cumulative proportion of 91.86%, where the first 24 principal components cover 89.79% of the total explained variance. Hence, holding on to our threshold, the first 25 dimensions will be kept. Then PNS was applied to the data, the data was projected onto 29 sub spheres, the relative explained variance $V(\ell)$ was as follows:

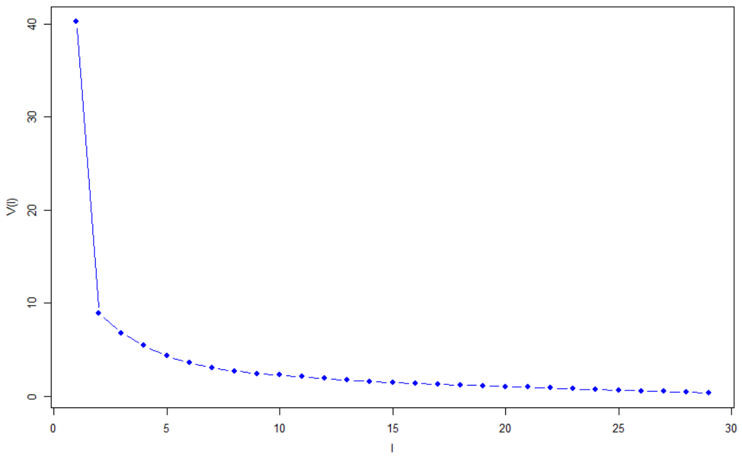


Figure 7: $V(\ell)$ for each subsphere

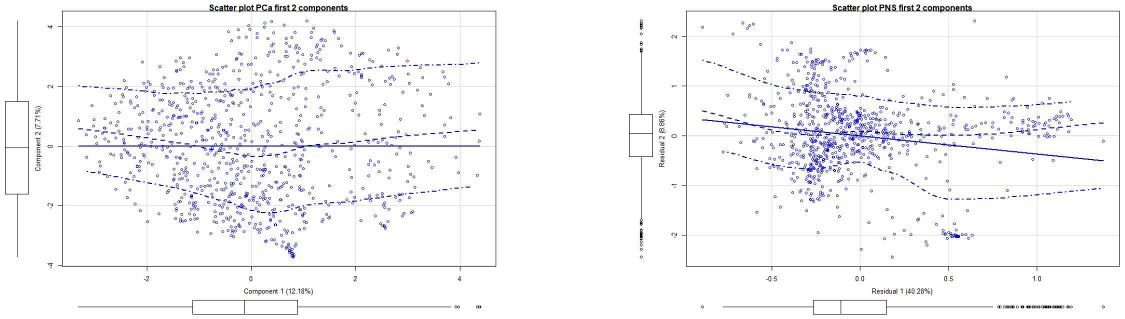
From the figure we can already see that the first components from PNS explain much more variability in the data then PCA, as the first two components seem to explain 49.14%, making a table for the first 10 components for comparison gives us:

Proportions of variance (%) for financial loss data										
Components	1	2	3	4	5	6	7	8	9	10
Principal Component Analysis	12.18	7.71	5.80	5.01	4.42	3.96	3.60	3.42	3.31	3.23
Principal Nested Spheres	40.28	8.86	6.78	5.42	4.31	3.57	3.02	2.71	2.42	2.25

Table 1: PCA and PNS comparison in explained variance

From table 1 we can see that PNS is clearly the better one, in terms of representing maximal variation using as few components as possible. As we expected, PCA using Euclidean distances does not respect the spherical structure of the data, hence it performs way worse then PNS in this case. Also, because the original PCA tries to find a rotation of the original data such that their covariance matrix is diagonal, and given the fact that we are in a multivariate extreme setting with a matrix that only contains small positive values after the Fréchet transformation, PCA does not capture quite lot variance in the first few components. In Cooley and Thibaud, 2016 and Drees and Sabourin, 2019 the original PCA has been modified to fit the extreme setting.

By choosing a threshold of 90% of total explained variance, only the first 17 components using PNS is chosen, this amount is notably lower then the 25 components from PCA. The scaled residuals from PNS can be seen as principal component scores, below are the scatterplots for the first 2 components of PCA and PNS :



(a) PCA scores for first 2 components

(b) PNS scores for first 2 components

Figure 8: Plot of scores for Financial portfolio data extreme values

From figure 8 we can see that the first 2 components from PNS give us much more information about the variability of the data then PCA. As the score points have a smaller randomness in their spread pattern for PNS.

Table 2 shows the total minimized distances from all observations to its center. PNS cleary outperforms PCA by having a smaller distance for a given k. Although toh PNS and PCA are outperformed by the original dataset on this criteria we can see that for larger k, the performance from PNS reduced dataset has a steeper increase then the original.

Table 2: Total minimized distance to cluster centers.

Total minimized distance	$W(A_n^5, S_n)$	$W(A_n^{10}, S_n)$
Financial portfolio loss	273.134	208.588
PCA reduced data	436.960	334.819
PNS reduced data	296.061	212.1578

5.2 Dietary intake data

The dietary data is again transformed with help of the empirical distribution function, and only the transformed observations with an Euclidean norm larger then 5% were kept. For the choice of the number of clusters we take a look at the elbow plot:

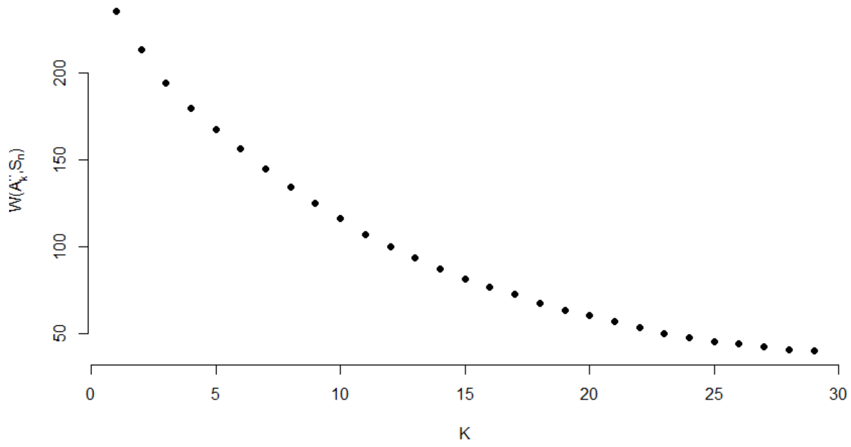


Figure 9: The value of the minimized mean distance $W(A_n^k, S_n)$ for different values of k in the financial portfolio loss data.

However, the elbow plot shows similarities to the plot from financial portfolio data, and hence the choice for k is again inconclusive. From the heatmaps shown in figures 10 and 11 it becomes clear that the number of clusters with only one large variable increases when k increases. This again, hints at a possible asymptotic independence of a majority of the nutrients. Clusters that are significant for several values of k can for example be identified as the clusters that are formed by sugar and carbs, or by saturated fat, sodium, total fat and calories. Furthermore, Vitamin B2, Vitamin B6, Vitamin B12 and niacin seem to go well together.

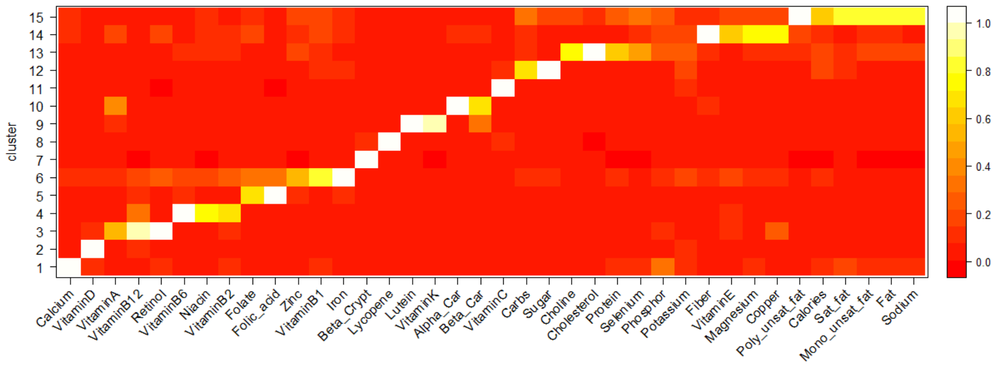


Figure 10: The k-means clustering result for dietary intake data for $k = 15$. Each row corresponds with to one of the ten estimated cluster centers, where values have been normalized as proportion of the largest variable. Which means that having a lighter color means that the corresponding component is relatively larger to all other components in the same center.

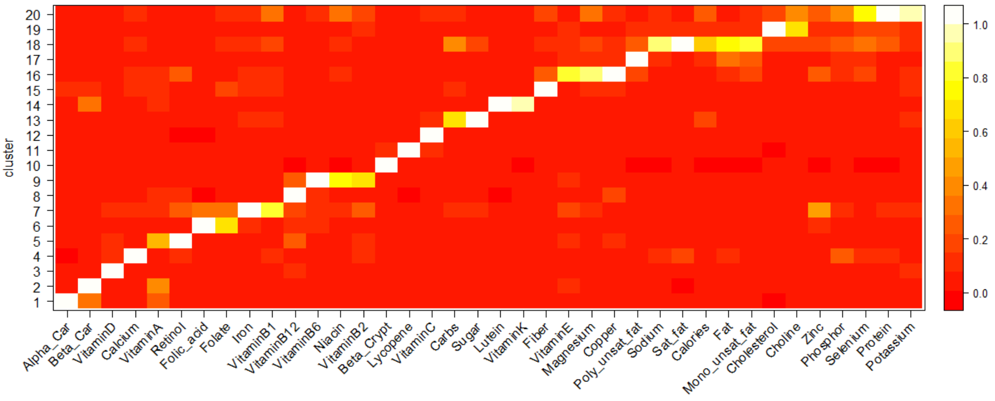
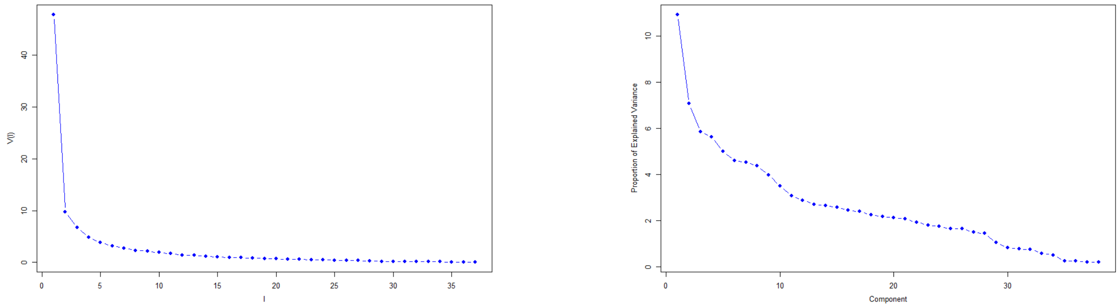


Figure 11: The k-means clustering result for dietary intake data for k = 20. Each row corresponds with to one of the ten estimated cluster centers, where values have been normalized as proportion of the largest variable. Which means that having a lighter color means that the corresponding component is relatively larger to all other components in the same center.

We again apply some dimension reduction methods for the dietary intake data, figure 12 shows that PSN is once more superior to PCA, capturing the most variability in the least amount of variables. By choosing a threshold of 90% total explained variance, only 14 dimensions are kept with PNS while we must keep 25 variables with PCA.



(a) First 38 principal nested spheres

(b) First 38 principal components

Figure 12: Plot of scores for Financial portfolio data extreme values

We again apply spherical k-means to the reduced dataset from PCa and PNs to compare the total minimized mean distance for $k = 15$ and $k = 20$

Table 3: Total minimized distance to cluster centers.

Total minimized distance	$W(A_n^{15}, S_n)$	$W(A_n^{20}, S_n)$
Dietary intake data	86.041	62.964
PCA reduced data	95.126	67.724
PNS reduced data	44.472	33.176

From table 3 we can see that PCA still does not really improve the clustering quality even for a large k, while the clustering results from PNS give a significantly lower distances from all observations to its cluster center compared to the other two datasets for a larger k.

6 Conclusion

In this paper, multiple methods to analyze extreme values from multivariate data through clustering and dimensionality reduction were investigated. Observations with a Euclidean norm that were considered large enough were used in the analysis. The dependencies between extreme observations were modelled by estimating a spectral measure through spherical k-means clustering. The methods used for dimensionality reduction were PCA and PNS.

The main theory behind extreme values analysis were explained in section 4.1 , in which the necessary conditions for measuring extremal dependencies were explained. In sections 4.2 the theory for modeling the spectral measure was given in form of the classical non-parametric max-linear model. Sections 4.3 explained the empirical estimation of the model by using a clustering approach. Furthermore, section 4.4 provided some dimensionality reduction techniques with respect to a Euclidean space and a sphere. Finally, in section 5 some real life data examples were given for illustrating the interpretation of cluster centers and the comparison between the dimensionality reduction methods.

From the results in section 5 it we diagnosed that there is a strong hint on asymptotic independence between the components in case there is only one relatively large component in the cluster. While asymptotic dependence was assumed when there were several large components in one cluster.

For dimensionality reduction, the number of principal components used for PCA and PNS are based on the total explained variance. In both datasets, PNS performed better then PCA. PNS captured more explained variance then PCA, while still having less variables then PCA.

For future interests, it is of interest to further develop PCA method, and optimize it in a extreme setting, and test it on datasets with more dimensions then we had in this paper.

References

Aslanertik, B. E., Erdem, S., & Kurt Gümüş, G. (2017). Extreme value theory in finance: A way to forecast unexpected circumstances. In H. Dinçer & Ü. Hacioğlu (Eds.), *Risk management, strategic thinking and leadership in the financial services industry : A proactive approach to strategic thinking* (pp. 177–190). Springer International Publishing. https://doi.org/10.1007/978-3-319-47172-3_12

Beirlant, J., Goegebeur, Y., Teugels, J., & Segers, J. (2004). *Tatistics of extremes: Theory and applications* [ISBN: 9780470012383]. John Wiley Sons, Ltd. https://doi.org/10.1002/0470012382

Bellman, R. (1957). *Dynamic programming* [ISBN: 069107951X, 9780691079516]. Princeton University Press.

Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics*, *9*, 383–418. https://doi.org/10.1214/15-EJS1002

Coles, S. (2001). *An introduction to statistical modeling of extreme valuess* [ISBN: 978-1-4471-3675-0]. Springer, London. https://doi.org/DOIhttps://doi-org.eur.idm.oclc.org/10.1007/978-1-4471-3675-0

Cooley, D., & Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, *106*(3), 587–604. https://doi.org/10.1093/biomet/asz028

Cooley, D., & Thibaud, E. (2016). Principal component decomposition and completely positive decomposition of dependence for multivariate extremes.

Dhillon, M. D., I.S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*, 143–175. https://doi.org/https://doi-org.eur.idm.oclc.org/10.1023/A:1007612920971

Drees, H., & Sabourin, A. (2019). Principal component analysis for multivariate extremes.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling extremal events for insurance and finance* [ISBN: 978-3-642-33483-2]. Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-33483-2

Engelke, S., & Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, *8*(1), 241–270. https://doi.org/10.1146/annurev-statistics-040620-041554

Fisher, R., & Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest members of a sample. *Proceedings of the Cambridge Philosophical Society*, *24*, 180–190. https://doi.org/http://dx.doi.org/10.1017/S0305004100015681

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathématique*, *6 (1)*, 93–116.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, *44 (3)*, 423–453. https://doi.org/https://doi-org.eur.idm.oclc.org/10.2307/1968974

Gumbel, E. J. (1958). *Statistics of extremes* [ISBN: 978-0-486-43604-3]. Mineola, NY: Dover.

Gumbel, E. (1935). "les valeurs extrêmes des distributions statistiques. *Annales de l'Institut Henri Poincaré*, *5 (2)*, 115–158.

Isabelle Guyon, A. E. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Janßen, A., & Wan, P. (2020). K-means clustering of extremes. *14*, 1211–1233. https://doi.org/https://doi.org/10.1214/20-EJS1689

JUNG, S., DRYDEN, I. L., & MARRON, J. S. (2012). Analysis of principal nested spheres. *Biometrika*, *99*(3), 551–568. http://www.jstor.org/stable/41720714

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297.

Mirkin, B. (1998). Mathematical classification and clustering: From how to what and why. in classification, data analysis, and data highways. *Springer*, 172–181. https://doi.org/https://doi-org.eur.idm.oclc.org/10.1007/978-3-642-72087-1_20

Mises, R. v. (1936). La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique*, *1*, 141–160.

Morton, I., & Bowers, J. (1996). Extreme value analysis in a multivariate offshore environment. *Applied Ocean Research*, *18(6)*, 303–317. https://doi.org/https://doi.org/10.1016/S0141-1187(97)00007-2

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *2(11)*, 559–572.

Resnick, S. I. (1987). *Extreme values, regular variation and point processes* [ISBN: 978-0-387-75953-1]. Springer-Verlag New York. https://doi.org/10.1007/978-0-387-75953-1

Appendix A : Rstudio codes

```
library(clustrd)
library(readxl)
library(BBmisc)
library(dplyr)
library(factoextra)
library(foreach)
library(doParallel)
library(lattice)
library(skmeans)
library(cluster)
library(seriation)
library(gplots)
###########################################################
#### import and construct the financial portfolio data ####
###########################################################

industry_portfolio <- read.csv("D:/value-averaged daily returns
of 30 industry portfolios.csv")
Y <- industry_portfolio[,-1] * (-1)
dates <- dat[,1]

###########################################################
#### transforming data with its empirical distribution ####
###########################################################

for(j in 1:ncol(Y)) {
  F <- ecdf(Y[,j]);
  nam <- paste("F", j, sep = "")
  assign(nam, F)
  for(i in 1:nrow(Y)) {       # for-loop over rows
    Y[i, j] <- 1/(1-F(Y[i,j]))
  }
}
```

```
################################################################################
#### Calculate the euclidean norms and keep observations with largest norms ####
################################################################################

e_norm <- function(x) sqrt(sum(x^2))
l<-length(Y[, 1])
norms <- vector(length = l)
threshold = round(0.05* l)

for(i in 1:nrow(Y)) {
  norms[i] <- e_norm(Y[i,])
}

maxima <- norms[order(-norms)][1:threshold]
indices_maxima <- match(maxima, norms)

Y_max <- Y[indices_maxima,]

#################################################################################
#### projection of the maxima on the unit sphere and apply spherical k-means ####
#################################################################################
Y_max_unit <- Y_max
for(i in 1:nrow(Y_max)){
  Y_max_unit[i,] = Y_max[i,] / e_norm(Y_max[i,])
}
Y_max_unit_matrix <- data.matrix(Y_max_unit[-1:-4,])

value<-rep(0,30)
for (k in 2:30){
  value[k]<-skmeans(Y_max_unit_matrix,k,method="pclust",control = list(nruns = 1000))$value
  print(k)
}
plot(value[2:30],ylab=expression(paste(W, "(", A[k]^n, ",", S[n], ")")), xlab="k")
plot(value[2:30],type="p", pch = 19, frame = FALSE, xlab="K",
     ylab=expression(paste(W, "(", A[k]^n, ",", S[n], ")")))
```

```
clus_Y <- skmeans(Y_max_unit_matrix, k = 5, method = "pclust", m = 1.2, weights = 1,
                  control = list(nruns = 1000, maxchains=100))

#####################################
#### plot event time over years ####
###################################
cluster.list <- rep(NA,length(clus_Y$cluster))
for (i in 1:k){
  cluster.list[clus_Y$cluster==col.ind[i]] <- i
}
par(mfrow=c(1,1))
par(mar=c(2,4,2,2))
plot(dates,cluster.list,pch=4,xlab=NULL,ylab='cluster')

##############################
#### plot cluster density ####
##############################
centroids <- clus_Y$prototypes
centroids <- apply(centroids,1,function(x){x/max(x)})
hm <- heatmap(centroids,scale='none')
row.ind <- hm$rowInd
centroids <- centroids[row.ind,]
col.ind <- order(apply(centers,2,which.max))
centers <- centers[,col.ind]
cat <- row.names(centroids)
data <- expand.grid(category=cat,dim=as.character(1:k))
data$Z <- as.vector(centers)
levelplot(Z ~ category*dim, data=data ,
col.regions =  heat.colors(100)[1:length(heat.colors(100))], main="",
xlab=NULL,ylab='cluster',
scales=list(y=(list(cex=1)), tck = c(1,0), x=list(rot=45,cex=1)))

##################################
#### Dimensionality reduction ####
##################################
library(shapes)
```

```
pns.out <- pns(t(Y_max_unit.ext))
proportion.pns <- pns.out[["percent"]]
x <- 1:29

plot(x, proportion.pns, type = "b", pch = 19, col = "blue", ylab = "V(l)", xlab = "l")



library(ggplot2)
pca.out <- prcomp(Y_max_unit, scale = TRUE, center = TRUE)
variance <- pca.out$sdev^2
proportion <- variance / sum(variance)
proportion <- proportion * 100

proportiondata <- as.data.frame(proportion)

x <- 1:30

plot(x, proportion, type = "b", pch = 19, col = "blue", ylab = "Proportion of Explained Vari

plot(pca.out$x[,1], pca.out$x[,2], main="Scatterplot PCA",
     xlab="Component 1", ylab="Component 2", pch=19)
abline(lm(pca.out$x[,2]~pca.out$x[,1]), col="red") # regression line (y~x)
lines(lowess(pca.out$x[,1],pca.out$x[,2]), col="blue") # lowess line (x,y)

clusters <- 5
pcaskmeans <- skmeans(pca.out$x[, 1:25],clusters,method="pclust",control = list(nruns = 100
pnsskmeans <- skmeans(t(pns.out$resmat[1:17,]),clusters,method="pclust",control = list(nruns



library(car)
scatterplot(pca.out$x[,1] ~ pca.out$x[,2]| type,  gropus = 5, data=stock.ext,
            xlab="Component 1 (12.18%)", ylab="Component 2 (7.71%)",
            main="Scatter plot PCa first 2 components",
            labels=row.names(stock.ext))
```

```
plot(t(pns.out$resmat[1,]), t(pns.out$resmat[2,]), main="Scatterplot PNS",
     xlab="Residual 1", ylab="Residuals 2", pch=19)
abline(lm(t(pns.out$resmat[2,])~t(pns.out$resmat[1,])), col="red") # regression line (y~x)
lines(lowess(t(pns.out$resmat[1,]),t(pns.out$resmat[2,])), col="blue") # lowess line (x,y)

scatterplot(pns.out$resmat[1,] ~ pns.out$resmat[2,],  data=stock.ext,
            xlab="Residual 1 (40.28%)", ylab="Residual 2 (8.86%)",
            main="Scatter plot PNS first 2 components",
            labels=row.names(stock.ext))
```