Analysis of Multivariate Extremes via Clustering Techniques

by

Shiyuan Deng

(Under the Direction of Shuyang Bai)

Abstract

This study examines the integration of multivariate extreme analysis within clustering techniques, specifically focusing on spherical k-means clustering and spherical k-principal component clustering. We propose an approach to estimate linear factor models using spherical clustering methods, enhancing order selection through a novel penalized silhouette method for optimal cluster number determination. This penalized silhouette method addresses limitations in traditional order selection by incorporating a penalty term, improving the accuracy of cluster identification in high-dimensional data.

Furthermore, we demonstrate the utility of sparse spherical k-principal component clustering in identifying groups of concomitant extremes, which is crucial in contexts where extreme values play a dominant role, such as in risk management or environmental modeling. This sparse clustering approach allows for efficient dimension reduction and identifies relevant factors while preserving the interpretability of extreme groupings.

Our findings suggest that the proposed spherical clustering techniques provide robust solutions for analyzing and grouping multivariate extremes, offering an effective framework for high-dimensional data where conventional clustering methods may fall short. By enhancing the interpretability and precision of cluster detection, this research contributes valuable insights to fields requiring accurate analysis of extreme values, supporting improved data-driven decision-making.

Index words: Multivariate Extremes, Spherical K-Means Clustering, Spherical k-Principal Component Clustering, Linear Factor Models, Penalized Silhouette Method, Cluster Order Selection, Sparse Clustering, Concomitant Extremes.

Analysis of Multivariate Extremes via Clustering Techniques

by

Shiyuan Deng

B.S., Beijing Institute of Technology, 07/2017, CHINA
M.S., University of Minnesota-Duluth, 2017

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

Doctor of Philosophy

Athens, Georgia

2024

Analysis of Multivariate Extremes via Clustering Techniques

by

Shiyuan Deng

Major Professor:	Shuyang Bai
Committee:	Ting Zhang
	T.N. Sriram
	Yuan Ke

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2024

Dedication

For my family, my past - The roots that shaped me and the stories that shaped me.

Acknowledgments

I would like to express my deepest gratitude to Shuyang Bai for their invaluable guidance and support throughout this work. I am also grateful to my colleague and collaborator, He Tang, whose constructive feedback and insightful discussions greatly enriched this research. Special thanks to my former advisor Yongcheng Qi, who brought me to this area and gave the chance for studying abroad. Finally, I extend my heartfelt appreciation to my parents, Shifang Xiao and Yanhong Deng, and my friend Jialin Yang for their unwavering encouragement and understanding throughout this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Extreme Value Analysis

During the development of statistical models, there are situations where people focus specifically on extreme events, such as floods, earthquakes, or financial crises. In these cases, rather than analyzing all the available observations, extreme value studies concentrate on the behavior of the most extreme values. This is because these rare occurrences can have the greatest impact. For example, in flood risk assessments, the maximum rainfall during a storm may serve as a critical indicator for predicting and managing flood disasters.

Extreme value theory is particularly useful in such cases because it allows us to model and understand the probabilities and magnitudes of these rare, extreme events. By focusing on the extremes, we gain insights into the potential severity of events that are not well-represented by average or typical observations, helping to make informed decisions about risk management and preparedness.

Extreme value theory in statistics has been developed over many decades and is now an important tool for understanding rare but significant events. It started with studying single variables and has since grown to cover multiple variables, with many practical applications across different fields.

In environmental science, extreme value analysis is used to predict the likelihood of extreme weather events, such as hurricanes, heatwaves, or floods, which are essential for risk assessment and disaster management. In finance, extreme value analysis helps model extreme market fluctuations, guiding risk management strategies and regulatory compliance. Engineering applications leverage extreme value analysis to design structures capable of withstanding rare stresses, such as those caused by earthquakes or high winds. Similarly, in public health, extreme value analysis aids in studying the spread of rare diseases or extreme exposure to hazardous substances. By focusing on the tails of distributions,

extreme value techniques provide valuable insights into the behavior of phenomena that standard statistical models often overlook, enabling more robust decision-making in the face of uncertainty.

1.1 Univariate Extreme Value Theory Recap

Univariate extreme value theory (EVT) and the statistical tools are well-established in dealing with rare events. The history of EVT dates back to the early 20th century, Fréchet, 1927 and Von Mises, 1936 studied the properties of extreme values of independent and identically distributed (IID) random variables. In the 1950s, Rényi, 1963 extended these ideas to the case of non-IID random variables, which provided the theoretical foundation for the development of multivariate EVT.

One fundamental result from the EVT is the Fisher–Tippett–Gnedenko theorem developed by the work of Fréchet, 1927, Fisher and Tippett, 1928, Von Mises, 1936, Falk and Marohn, 1993 and Gnedenko, 1943. Here, we provided a short review of the results: Suppose $X_1, X_2, \cdots, X_n$ are IID random variables with cumulative distribution function F. Suppose there exist two real number sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that the following normalized sample maximum converges in distribution to a non-degenerate distribution function G:

$$\lim_{n\to\infty} P\left(\frac{\max\{X_1, X_2, \cdots, X_n\} - b_n}{a_n} \leq x\right) = G(x), \quad x \in \mathbb{R}. \tag{1.1}$$

Then the limit distribution of G necessarily belongs to one of 3 possible families of distributions, the Gumble, the Frechét, and the Weibull.

These three families can be combined into a single family known as the generalized extreme value distribution family given by the following formula:

$$G(x) = \exp\left(-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right), \quad 1+\gamma\frac{x-\mu}{\sigma} > 0, \tag{1.2}$$

where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter and $\gamma \in \mathbb{R}$; when $\gamma = 0$, the expression $\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}$ is understood as $\exp\left(-\frac{x-\mu}{\sigma}\right)$ through the limit as $\gamma \to 0$. The parameter γ, known as the *shape parameter*, or the *extreme value index*, plays a critical role in controlling the properties of G. The Gumble, the Frechét, and the Weibull distributions correspond to the

parameter ranges $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$ respectively, whose cumulative distribution functions are displayed below:

- **Type I or *Gumbel* extreme value distribution**, case $\gamma = 0$, for all $x \in (-\infty, +\infty)$:

$$F(x; \mu, \sigma, 0) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right).$$

- **Type II or *Fréchet* extreme value distribution**, case $\gamma > 0$, for all $x \in \left(\mu - \frac{\sigma}{\gamma}, +\infty\right)$:

$$\text{Let } \alpha \equiv \frac{1}{\gamma} > 0 \quad \text{and} \quad y \equiv 1 + \frac{\gamma}{\sigma}(x-\mu).$$

$$F(x; \mu, \sigma, \gamma) = \begin{cases} 0, & y \leq 0 \quad \text{or equivalently } x \leq \mu - \frac{\sigma}{\gamma}, \\ \exp\left(-\frac{1}{y^\alpha}\right), & y > 0 \quad \text{or equivalently } x > \mu - \frac{\sigma}{\gamma}. \end{cases}$$

- **Type III or *reversed Weibull* extreme value distribution**, case $\gamma < 0$, for all $x \in \left(-\infty, \mu + \frac{\sigma}{|\gamma|}\right)$:

$$\text{Let } \alpha \equiv -\frac{1}{\gamma} > 0 \quad \text{and} \quad y \equiv 1 - \frac{|\gamma|}{\sigma}(x-\mu).$$

$$F(x; \mu, \sigma, \gamma) = \begin{cases} \exp\left(-y^\alpha\right), & y > 0 \quad \text{or equivalently } x < \mu + \frac{\sigma}{|\gamma|}, \\ 1, & y \leq 0 \quad \text{or equivalently } x \geq \mu + \frac{\sigma}{|\gamma|}. \end{cases}$$

Similar to the central limit theorem which indicates sum-stability of Gaussian, the Fisher–Tippett–Gnedenko theorem also indicates the $G(x)$ distribution is max-stable, and the max-stable property a central role in extreme value theory as they describe the possible limiting distributions for normalized maxima of random variables. They provide a theoretical basis for modeling and understanding extreme events in a wide range of fields.

1.1.1 Approaches to Identify and Model the Extrema

There are several methods for modeling extreme values, with the most common approaches being the Peaks Over Threshold (POT) and Block Maxima (BM) methods. To illustrate these two methods, we use a simulated example.

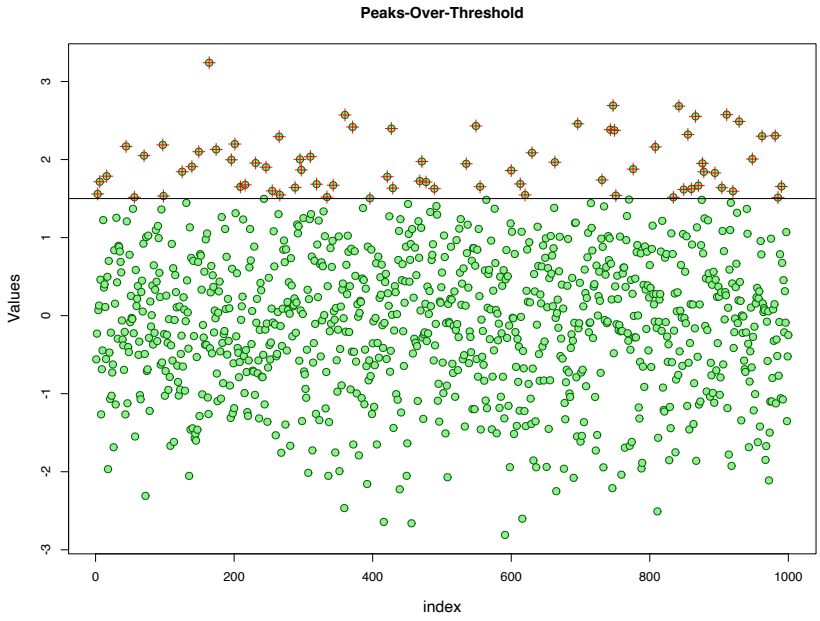


Figure 1.1: POT Example

A random sample of 1000 observations is generated from the standard normal distribution, as shown in Figures 1.1 and 1.2.

Peaks Over Threshold

In the POT approach, illustrated in Figure 1.1, extreme values are identified by focusing on data points that exceed a specified high threshold (value = 1.5 in Figure 1.1). This method captures all values above the threshold, rather than only the single maximum within a given period, allowing for more detailed modeling of extreme events. The excesses above the threshold are typically modeled using the Generalized Pareto Distribution (GPD), which offers flexibility in characterizing the tail behavior of the data. POT is particularly effective in scenarios where data is sparse but the focus is on extreme deviations, such as in insurance claims or financial losses.

Block Maxima

The Block Maxima approach divides the data into equally sized blocks (e.g., days, months, or years) and selects the maximum value from each block as shown in Figure 1.2. These maxima are then modeled using the Generalized Ex-

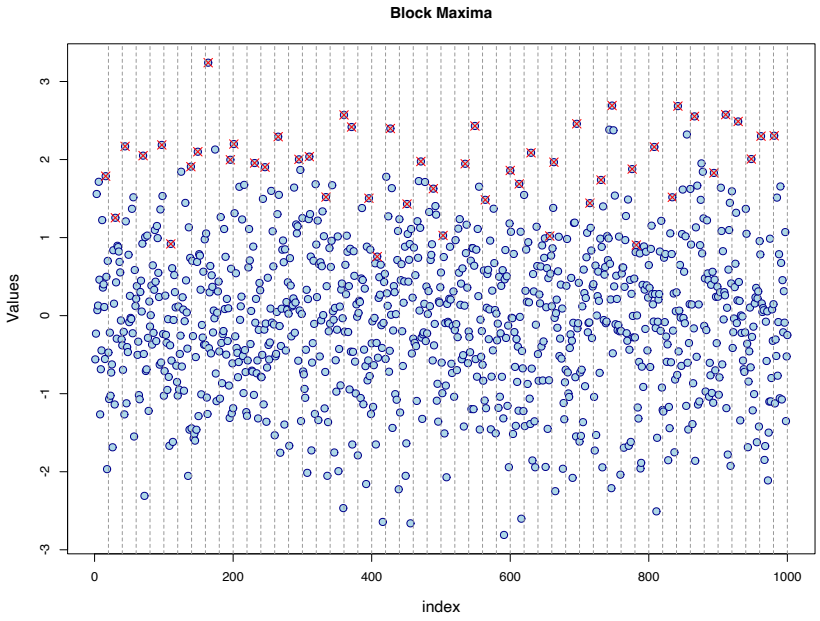


Figure 1.2: BM Example

treme Value (GEV) distribution, which combines several distributions (Gumbel, Fréchet, and Weibull) to describe the behavior of extremes. This method is suitable for long-term analysis, such as studying yearly maximum temperatures or annual flood levels.

The Peaks Over Threshold (POT) and Block Maxima (BM) methods are two fundamental approaches in extreme value theory, each with its distinct advantages. Block Maxima divides data into blocks (e.g., years, months) and models the maximum value within each block using the Generalized Extreme Value (GEV) distribution, which simplifies analysis but may discard relevant extreme data below the maximum. On the other hand, the POT method focuses on values exceeding a high threshold, capturing more extreme data points and providing more efficient use of data through the Generalized Pareto Distribution (GPD). While BM is simpler to implement, POT is often preferred for more precise modeling of the tail behavior when more extreme data is available.

In addition to the frequentist approaches, Bayesian methods take the prior information into consideration to model the structure of the extreme values; for example, Bottolo et al., 2003 view exceedances over a given threshold as a

Poisson point process and used a Bayesian method to develop a hierarchical mixture prior.

EVT has been applied in a wide range of fields, including hydrology, finance, engineering, and climate science, among others. In hydrology, EVT is used to model the distribution of extreme rainfall and floods (Engeland et al., 2004). In finance, EVT is used to model the distribution of extreme stock market returns and losses (Hussain and Li, 2015). In engineering, EVT is used to model the distribution of extreme loads on structures (Chen, 2014).

1.2 Multivariate Extreme Value Theory

Multivariate Extreme Value Theory (MEVT) extends the concepts of univariate EVT to situations where multiple variables are involved. It focuses on understanding the joint behavior of extremes in multivariate datasets, which is particularly important in applications where extreme events often occur simultaneously or are interconnected across different variables, such as in finance, meteorology, or engineering.

In finance, MEVT is critical for modeling the simultaneous occurrence of extreme losses across multiple assets or markets, enabling better risk assessment and portfolio management strategies. In meteorology, it is used to analyze the joint behavior of extreme weather phenomena, such as the combined impact of heavy rainfall and strong winds during storms, providing insights for disaster preparedness and infrastructure planning. Engineering applications leverage MEVT to evaluate the combined extremes of load and stress on structures, ensuring their resilience under rare but critical conditions. By capturing the dependencies between extreme events across variables, MEVT offers a more comprehensive framework for understanding and mitigating the risks associated with multivariate extremes in complex systems.

As a generalization of (1.1), Pickands–Balkema–De Haan theoremBalkema and De Haan, 1974 provides the theoretical foundation for modeling the joint distribution of extreme events in multiple dimensions.

Let $(X_1^i, \cdots, X_d^i), i \in \mathbb{N}$ be IID copies of random vector $\mathbf{X} = (X_1, \cdots, X_d)$ assumed to follow a continuous joint distribution for simplicity. Assume that there exist sequences of constants $a_j^n > 0,\ b_j^n \in \mathbb{R}, 1 \leq j \leq d, n \in \mathbb{N}$ and a joint distribution function G, such that

$$\lim_{n\to\infty} P\left(\frac{\max\limits_{i=1,\cdots,n} X_1^i - b_1^i}{a_1^n} \le x_1, \cdots, \frac{\max\limits_{i=1,\cdots,n} X_d^i - b_d^i}{a_d^n} \le x_d\right)$$
$$= G(x_1, \cdots, x_d), \tag{1.3}$$

for all the continuity points $(x_1, \cdots, x_d)$ of G. The analysis of multivariate extremes typically consists of the modeling of marginal tails and the modeling of extremal dependence.

Accordingly, the convergence in (1.3) has two parts. First, all marginal $G_j,\ 1 \le j \le d$, of G are of the form

$$G_j(x) = \exp\left(-\left(1 + \gamma_j \frac{x-\mu_j}{\sigma_j}\right)^{-1/\gamma_j}\right),\quad 1 + \gamma_j \frac{x-\mu_j}{\sigma_j} > 0, \tag{1.4}$$

where $\mu_j \in \mathbb{R}$, $\sigma > 0$ and $\gamma_j \in \mathbb{R}$, which means that each of the marginal distribution is a univariate extreme distribution as in (1.2). Second, let F_j be the (continuous) marginal distribution function of $X_j,\ j = 1, \cdots, d$. Then the convergence (1.3) holds iff the marginally standardized vector

$$\mathbf{Y} = (Y_1, \ldots, Y_d) = \left(\frac{1}{1 - F_1(X_1)}, \cdots, \frac{1}{1 - F_d(X_d)}\right) \tag{1.5}$$

satisfies

$$\lim_{y\to\infty} P\left(\frac{\mathbf{Y}}{\|\mathbf{Y}\|} \in B \mid \|\mathbf{Y}\| > y\right) = S(B), \tag{1.6}$$

with a probability measure S on $\mathbb{S}_+^{d-1} := \left\{\mathbf{x} \in [0,\infty)^d : \|\mathbf{x}\| = 1\right\}$ and any S-continuity-Borel set $B \subset \mathbb{S}_+^{d-1}$. In principle, the norm $\|\cdot\|$ can be chosen as a fixed arbitrary norm. Throughout this thesis, we shall work with the L_2 norm $\|(x_1, \ldots, x_d)\| = (x_1^2 + \ldots + x_d^2)^{1/2}$, $(x_1, \ldots, x_d) \in \mathbb{R}^d$, due to its mathematical convenience.

Definition 1 *The measure S in* (1.6) *is called the* spectral measure.

The spectral measure S describes the limit behavior of the extremal dependence of X: a high concentration of the measure S on certain region $B \subset S$ means a high chance of observing an extremal concurrence of $(Y_1, \ldots, Y_d)$, a marginally normalized version of $(X_1, \ldots, X_d)$, to appear in the directions contained by B.

With the standardization in (1.5), the distribution of the normalized componentwise maxima also converges to an multivariate extreme value distribution with a standard Fréchet margin ($\gamma = 1$ in (1.2)):

$$\begin{aligned}&\lim_{n\to\infty} P\left(\frac{\max_{i=1,\cdots,n} Y_1^i}{n} \leq x_1, \cdots, \frac{\max_{i=1,\cdots,n} Y_d^i}{n} \leq x_d\right)\\ &= G_0(x_1, \cdots, x_d),\end{aligned} \tag{1.7}$$

where $(Y_1^i, \cdots, Y_d^i), i \in \mathbb{N}$ are IID copies of $\mathbf{Y}$. The joint distribution G_0 in (1.7) can be expressed in the form:

$$\begin{aligned}&G_0(x_1, \cdots, x_d)\\ &= \exp\left(-\nu\left\{(u_1, \cdots, u_d) \in [0,\infty)^d \setminus \{\mathbf{0}\} : \exists j : u_j > x_j\right\}\right)\end{aligned} \tag{1.8}$$

for all $(x_1, \cdots, x_d) \in [0,\infty)^d \setminus \{\mathbf{0}\}$, for some σ-finite infinite measure ν on $[0,\infty)^d \setminus \{\mathbf{0}\}$.

Definition 2 *The measure ν characterized by* (1.8) *is known as the* exponent measure.

On the other hand, the exponent measure ν arises from the limit as follows

$$\lim_{t\to\infty} tP(\mathbf{Y}/t \in A) = \nu(A) \tag{1.9}$$

where $\mathbf{Y}$ is defined in (1.3) and A is any $\nu-$continuous Borel set in $[0,\infty)^d \setminus \{\mathbf{0}\}$ away from the origin (see, e.g., Engelke and Ivanovs, 2021). The *exponent measure* and the *spectral measure* are related through the following relation: for any Borel $B \subset \mathbb{S}_+^{d-1}$,

$$\nu\left\{\mathbf{u} \in [0,\infty)^d \setminus \{\mathbf{0}\} : \mathbf{u}/\left\|\mathbf{u}\right\| \in B, \left\|\mathbf{u}\right\| > y\right\} = cy^{-1}S(B), \quad y > 0, \tag{1.10}$$

for some constant $c > 0$.

To sum up, the study of multivariate extremal dependence is essentially a study of the structure of the *spectral measure* S.

For the marginally standardized observations (cf. (1.5)), a non-parametric estimator of the spectral measure S arising from the empirical version of (1.6) was proposed by Einmahl et al., 2001. Specifically, it is constructed by first

replacing all the marginal cumulative distributions $F_j,\ 1 \leq j \leq d$ of X in (1.5) by the empirical cumulative distribution functions

$$F_{j,n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{x_j^i < x\}}, \quad x \in \mathbb{R}.$$

Accordingly, the empirical version of transformed observation $\mathbf{Y}$ is

$$\hat{\mathbf{Y}}^i = (\hat{Y}_1^i, \cdots, \hat{Y}_d^i) \text{ with } \hat{Y}_j^i := (1 - F_{j,n}(X_j^i))^{-1}. \tag{1.11}$$

Thus, an estimator of spectral measure S is naturally given by

$$\hat{S}_n(B) := \frac{\sum_{i=1}^{n} \mathbb{I}_{\left\{\|\hat{\mathbf{Y}}_i\| \geq \frac{n}{l_n}, \frac{\hat{\mathbf{Y}}_i}{\|\hat{\mathbf{Y}}_i\|} \in B\right\}}}{\sum_{i=1}^{n} \mathbb{I}_{\left\{\|\hat{\mathbf{Y}}_i\| \geq \frac{n}{l_n}\right\}}}, \tag{1.12}$$

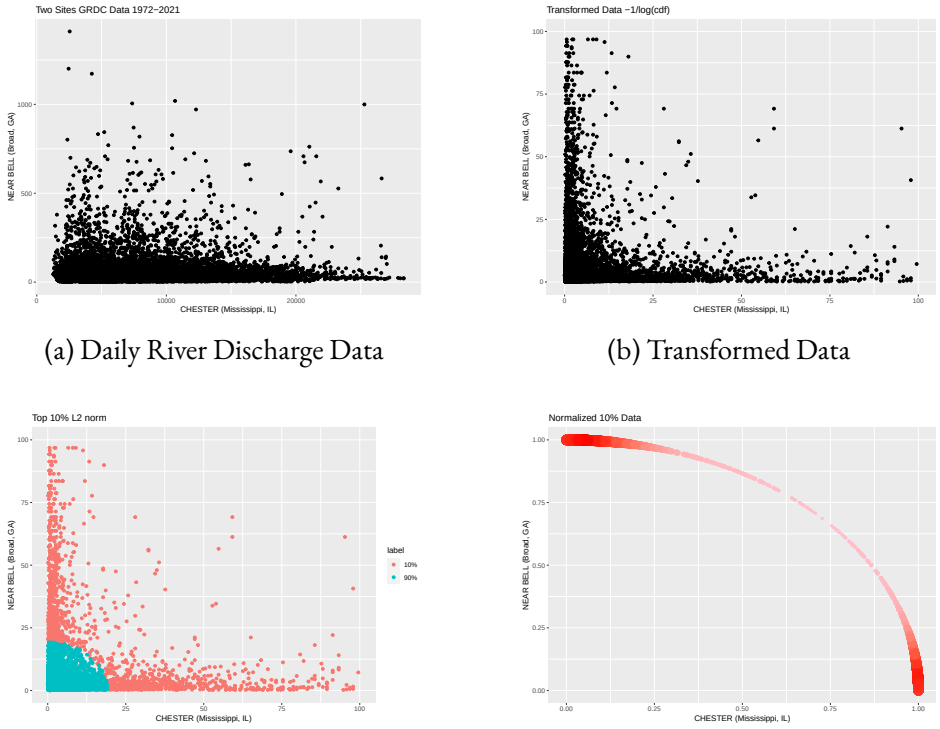
where B is a Borel subset of $\mathbb{S}_+^{d-1}$, and the threshold $l_n \in \mathbb{N}$ decides the number of extremal observations used for the estimator.

The choice of l_n in (1.12) is a issue has been discussed a lot. It is a classic topic in extreme value analysis which involves a bias-variance tradeoff. For instance, Bader et al., 2018 developed an efficient technique for threshold selection based on the Anderson–Darling test. Besides, Wan and Davis, 2019 proposed a procedure for selecting the threshold in modeling multivariate heavy-tailed data by testing the independence of the radial and angular components using distance covariance. This approach aims to improve the estimation of tail dependence by incorporating a subsampling scheme, reducing computational demands, and demonstrating its effectiveness on both simulated and real data.

1.2.1 An Illustration of Extremal Spectral Sample

To be clear about the extremal spectral sample, we used the daily river discharge data from two sites: "NEAR BELL" in Broad, GA and "CHESTER" in Mississipi, IL. The river discharge data are related to the daily discharge rate of rivers in North America sourced from the Global Runoff Data Center (German Federal Institute of Hydrology, n.d.). Scatter plots are drawn to illustrate the extremal spectral sample. The dataset comprises 16,386 daily records of discharge values from the two stations spanning the period from December 1, 1976, to October 11, 2021.

Figure 1.3a displays scatter plots of daily discharge data from two sites, highlighting some large values that may be strongly associated with flood events.



(a) Daily River Discharge Data

(b) Transformed Data

(c) 10% Data with Top L2 Norm

(d) Projection of 10% of Data

Figure 1.3: An Illustration of Extremal Spectral Sample

Detecting patterns in extreme values between the two sites is challenging, so we apply a transformation $-1/\log F_i$, $i = 1, 2$ where $F_i$is the empirical cumulative distribution function (CDF) of site i (see Figure 1.3b). To focus on extreme behaviors, we identify the 10% of records with the largest $L2$ norms, shown as red points in Figure By projecting the selected data onto the unit sphere, we visualize the extreme behaviors, or spectral measure, as seen in Figure 1.3c. Given that one site is located in Georgia and the other in Illinois, the unit sphere projection shows mass near the endpoints, indicating that extreme daily discharge events between these geographically separated sites are largely independent.

1.2.2 An Illustration of Convergence Towards Discrete Spectral Measure

To clarify the convergence toward a discrete spectral measure, we provide a simulation example illustrated in Figure 1.4. The 10,000 data points are generated

using the max-linear model:

$$\mathbb{X} = (X_1, X_2, X_3) = \left(\max_{i=1,\dots,3} b_1^i Z_i, \dots, \max_{i=1,\dots,k} b_3^i Z_i \right),$$

where $b_1 = (1, 0, 0)$, $b_2 = (0, 1, 0)$, and $b_3 = (0, 0, 1)$. Here, Z_i are independent standard Fréchet random variables. Details of the max-linear model will be discussed in the following sections.

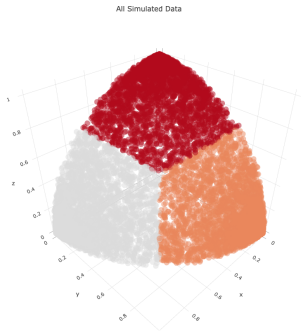
In Figure 1.4a, which displays all simulated data points on the unit sphere, detecting patterns in the data is challenging. By examining only the top 50% of data points based on the L_2 norm, as shown in Figure 1.4b, we can see that relatively large values are concentrated around the three vertices $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. When focusing on just the top 10% of data points (Figure 1.4c) with the greatest L_2 norm, it becomes even clearer that observations are clustered around these three vertices. These plots demonstrate the convergence of the spectral measure. As previously discussed, the choice of proportion for extreme values—denoted by l_n in 1.12—raises important considerations in analyzing spectral measures.

Studying the convergence toward a discrete spectral measure is essential in understanding the asymptotic behavior of extreme values in multivariate models, especially in fields that rely on modeling extreme events, such as finance, environmental science, and insurance. In max-linear models, as in our example, the data tend to concentrate near specific vertices of the unit sphere. This clustering behavior, as seen in the simulation, highlights a form of dependency structure among the extreme values, which provides insight into the underlying risk or interaction patterns between variables in high-dimensional spaces.

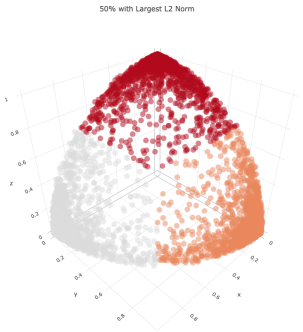
When convergence toward a discrete spectral measure occurs, it indicates that a small subset of directions or patterns dominates the extreme behavior of the system. This property enables simplification in modeling since it allows us to represent complex dependencies through a discrete spectral measure, capturing the essential behavior of extremes without the need to model every interaction in detail. For instance, in financial portfolios, it can reveal that only a few assets might drive extreme risks, which is valuable for risk management and mitigation.

Furthermore, understanding this convergence is critical when selecting appropriate thresholds and determining the proportion of extreme values to analyze. Different choices of thresholds can impact the interpretation of the spectral measure and, consequently, the understanding of the extremal dependence structure. As discussed, the choice of this threshold, denoted as l_n , directly influences how we assess convergence and measure stability. Hence, examining the

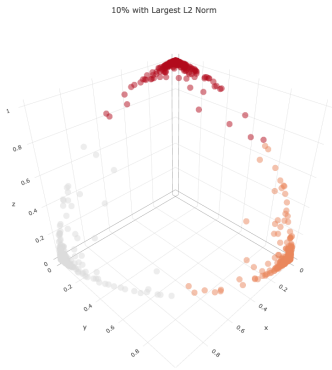
convergence properties enables researchers to make informed decisions regarding threshold selection, thereby improving the robustness and interpretability of the spectral measures in practical applications.



(a) All Simulated Data



(b) 50% Simulated Data



(c) 10% Simulated Data

Figure 1.4: Example of Convergence Towards Discrete Spectral Measure

1.2.3 A Three-dimensional Extreme Value Model Example

Here, we address a three-dimensional real case study. This study explores the joint occurrence of extreme daily river discharge recodes of three different locations. The data are obtained from the Global Runoff Data Centre German Federal Institute of Hydrology, n.d. The dataset comprises 16,386 daily records of discharge values from 13 stations spanning the period from October 1, 1976, to October 11, 2021. These 3 stations, shown in Figure 1.5, are positioned along 2 rivers in America: Willamette River, and Mississippi River.

combining generalized extreme value (GEV) distributions and a Gumbel copula. Each location's annual maximum wind speeds are modeled separately

Figure 1.5: Three Sites Location

using GEV distributions, capturing the statistical properties of extreme values, including their location, scale, and tail behavior. The GEV cumulative distribution functions (CDFs) transform these wind speeds into a uniform scale, facilitating the use of a Gumbel copula to capture the dependency structure between them. The Gumbel copula, particularly suited for modeling joint occurrences of high values, emphasizes the relationship between extreme wind events at the different sites, allowing for the estimation of joint probabilities of simultaneous extreme occurrences. A 3D visualization of the joint distribution illustrates these probabilities, revealing the likelihood of concurrent extreme wind conditions at all three locations.

Figure 1.6a displays the raw river discharge data from the three sites. The data is dense, with some clustering that likely corresponds to the discharge values recorded at the three locations. Due to the proximity of two sites, "PORTLAND" and "SALEM," their data may exhibit significant overlap or similar patterns, while the third site, "CHESTER," appears distinct, likely due to its geographic and hydrological differences. To better understand the data patterns, a transformation (1.11) is applied, as shown in Figure 1.6b. However, due to scaling issues, it remains challenging to discern the empirical patterns of the spectral measure.

To address this, the transformed data is normalized, as illustrated in Figures 1.6c, 1.6d, 1.6e, and 1.6f. Normalization enhances the clarity of the data patterns, making it easier to identify trends. Based on the normalized data, it becomes evident that the "PORTLAND" and "SALEM" sites share a similar trend, while the "CHESTER" site demonstrates distinct characteristics. By focusing on the "large" values—specifically the top 50%, 10%, and 1% of the data with the greatest L_2 norm—the convergence of the spectral measure becomes increasingly

clear. This approach highlights the dependence structure among the sites and emphasizes the divergence of "CHESTER" from the other two locations.

1.3 Extreme Value Statistics

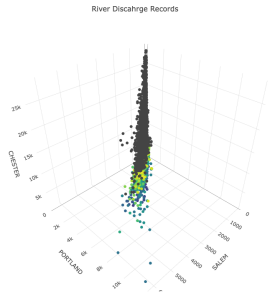
The sparsity property in multidimensional extreme value statistics plays a critical role, as it reflects the rare and isolated occurrences of extreme events across multiple dimensions. This sparsity often leads to intricate dependence structures, where the relationships between extreme values in different dimensions are neither straightforward nor uniformly distributed, creating a complex web of dependencies. To model these dependencies effectively, extreme value statistics leverages sophisticated methods that are adaptable to high-dimensional data. This adaptability has driven its applications across diverse domains, such as environmental sciences for assessing risks like floods or heatwaves, finance for extreme market movements, and engineering for infrastructure reliability under stress.

The integration of extreme value statistics with graphical models, machine learning, and causality has enriched both the methodological and application domains of this field. Graphical models, which represent variables as nodes and dependencies as edges in a network structure, allow researchers to visualize and quantify complex relationships among extremes in high-dimensional data. This is particularly beneficial when exploring sparsity-driven structures, as graphical models enable the simplification of complex dependency webs inherent in multivariate extremes. For example, in climate science, extremal graphical models can help identify joint patterns in temperature or precipitation extremes across spatially distributed locations, revealing interactions that inform regional risk assessments.

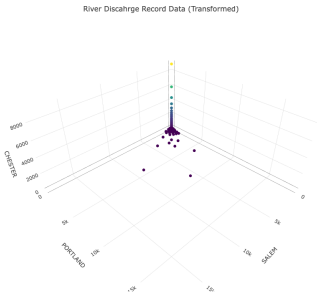
In machine learning, extreme value statistics has found application in training algorithms that must handle rare events, such as in predictive maintenance and anomaly detection. Neural networks and other machine learning models can leverage extreme value theory to enhance their predictive power for rare events, making it possible to forecast extremes with greater accuracy. For instance, algorithms trained on historical financial market data can incorporate extreme value distributions to better anticipate and prepare for market crashes or sudden financial anomalies, enhancing risk management systems. Additionally, generative models and other machine learning approaches facilitate data augmentation in scenarios with limited extreme observations, broadening the scope of potential applications.

Causality analysis has also benefited from extreme value statistics, particularly in domains where understanding causal drivers of extremes is essential. By combining causal inference techniques with extreme value modeling, researchers can distinguish between mere correlations and potential causal links in rare event occurrences. This approach is highly relevant in public health, where identifying the causative factors behind disease outbreaks can improve preventive measures. Similarly, in environmental sciences, determining causal factors for extreme weather events, such as links between greenhouse gas emissions and heatwaves, can aid in developing more effective climate policies.

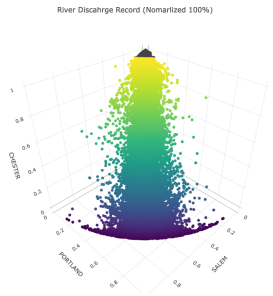
Together, these interdisciplinary applications of extreme value statistics underscore its versatility and impact across fields. By drawing on advancements in graphical models, machine learning, and causality, extreme value statistics not only enhances our understanding of dependencies in extreme events but also provides robust tools for predicting and managing these events across various sectors.



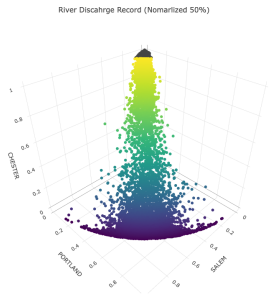
(a) Daily River Discharge Data



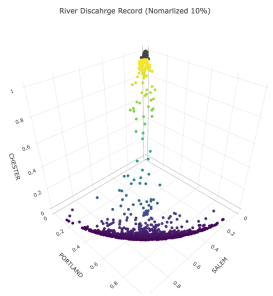
(b) Transformed Data



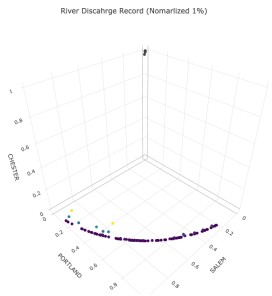
(c) All Normalized Data



(d) 50% Normalized Data



(e) 10% Normalized Data



(f) 1% Normalized Data

Figure 1.6: An Illustration of Extremal Spectral Sample

Chapter 2

Factor Models and Spherical Clustering Techniques

2.1 Linear Factor Models

Factor models are statistical tools designed to describe the relationships among a large number of variables by summarizing them in terms of a smaller set of unobserved latent variables, known as factors. These models provide a parsimonious representation of the data, making them invaluable for understanding underlying structures and simplifying complex systems.

In the context of extreme value theory (EVT), factor models are particularly useful for reducing dimensionality in multivariate extremes. By identifying latent factors that drive the dependence structure of extreme events, these models facilitate the analysis of high-dimensional data where extreme observations are rare and often spatially or temporally dependent. This makes them a powerful tool for studying phenomena such as heatwaves, financial crashes, or widespread insurance claims due to natural disasters.

Beyond EVT, factor models are widely applied in finance, where they are used to model the relationships between asset returns and macroeconomic factors or market-specific factors (e.g., in portfolio optimization or risk modeling). In economics, factor models help identify common trends or shocks affecting multiple economic indicators, such as GDP, inflation, and unemployment rates.

In risk management, factor models play a central role in identifying and quantifying risks associated with various factors, enabling firms to better manage exposure to market, credit, or operational risks. They are also utilized in

climatology to study large-scale atmospheric and oceanic processes, such as identifying dominant patterns in global temperature changes or rainfall extremes.

Other fields where factor models are applied include psychometrics (to measure latent traits such as intelligence or personality), bioinformatics (to uncover latent structures in genetic data), and marketing (to understand consumer preferences and segment markets). Their versatility and ability to handle high-dimensional data make factor models essential across numerous domains where reducing complexity while retaining interpretability is a priority.

Two common approaches to factor models in extreme value theory are the max-linear model and the sum-linear model. Both aim to represent the dependence structure of extreme values in a multivariate setting but do so in different ways.

The linear factor model has broad applications across fields due to its ability to capture the influence of underlying factors on observable outcomes. In finance, it explains asset returns through models like the Fama-French model. In econometrics, it assesses relationships among economic indicators. In machine learning, it underlies dimensionality reduction techniques like PCA. Additionally, variants like the max-linear model are used in risk management and engineering to analyze systems dominated by extreme factors. This flexibility makes the linear factor model a powerful tool for analyzing both collective and dominant factor impacts in diverse domains.

In our thesis, we examined two classical linear factor models: the max-linear and sum-linear models. Both models exhibit similar extreme behaviors and serve as foundational, straightforward tools for exploring the application of clustering methods in the subsequent sections.

2.1.1 Max-linear Model

Over the past decades, the construction of statistical models of multivariate extremes has been explored and discussed by researchers. One of the simplest models is the max-linear model. It was first introduced by Coles and Tawn in a series of papers in the 1990s (See. e.g., S. G. Coles and Tawn, 1991, S. G. Coles and Tawn, 1994, and S. Coles et al., 1999). Max-linear models are a class of models used to analyze multivariate data in which each variable is modeled as a max-linear combination of the values of some underlying latent variables. This type of model has found applications in fields such as environmental science, finance, and genetics, among others.

Model Formulation

One form of the max-linear model (Janßen and Wan, 2020a) can be represented as:

$$\mathbf{X} = (X_1, \cdots, X_d) = \left(\max_{i=1,\cdots,k} b_1^i Z_i, \cdots, \max_{i=1,\cdots,k} b_d^i Z_i \right) \tag{2.1}$$

where $\mathbf{b}_i = (b_i^1, \cdots, b_i^d) \in [0, \infty)^d, i = 1, \cdots, k$ are k different factors and $Z_1, \cdots, Z_k$ are IID random variables with the same heavy-tailed distribution, for instance, with the standard Fréchet distribution being a most common choice. We also impose an assumption that

$$\sum_{i=1}^{k} b_j^i = 1 \text{ for all } j = 1, \cdots, d, \tag{2.2}$$

so that the marginal distributions of $\mathbf{X}$ have the same rate of tail decay, e.g., they are each standard Fréchet if Z_i's are assumed so.

Application and challenages

Max-linear models have several advantages, including their flexibility in approximating complex extremal dependencies while maintaining simplicity and interpretability. A notable recent development of the max-linear model includes the studies on Bayesian networks, namely, directed acyclic graphical models, based on max-linear structural equations (See. e.g., Gissibl and Klüppelberg, 2018; Klüppelberg and Lauritzen, 2019).

However, the parameter estimation problem of max-linear models has been considered challenging mainly due to the fact that the likelihood is not available because of the lack of density for the spectral measure, which sets its parameter estimation problem for max-linear models apart from traditional likelihood-based parameter estimation procedures.

There have been several explorations on parameter estimation of max-linear models. Einmahl et al., 2012 and Einmahl et al., 2016 proposed using an M-estimator to minimize the distance between a vector of weighted integrals of the tail dependence function and their empirical counterparts. Yuen and Stoev, 2014 also introduced an M-estimation framework for max-stable models by utilizing the continuous ranked probability score (CRPS) of multivariate cumulative distribution functions. Recently, Janßen and Wan, 2020a discovered a connection between max-linear model estimation and spherical k-means clustering of extreme values.

Although methods of parameter estimation for the max-linear model have been considered as above, none of the methods have addressed model order selection, namely, the selection of k in (2.1). In this thesis, we first follow the work by Janßen and Wan, 2020a to relate the procedure of parameter estimation for max-linear models to a clustering process. Then the model selection problem is converted to a problem of selection of the number of clusters.

2.1.2 Sum-linear Model

The sum-linear model has developed as a fundamental approach for analyzing the additive effects of multiple factors on observable outcomes. It emerged from statistical methods to decompose complex systems into simpler components, allowing researchers to quantify how each factor contributes to the overall variability of a response variable. Over time, this model became integral in fields like finance (e.g., asset pricing models such as the Fama-French model), econometrics (to analyze economic indicators based on underlying factors), and psychology (for factor analysis in survey data). In machine learning, sum-linear models underpin dimensionality reduction techniques like Principal Component Analysis (PCA), helping to uncover latent structures and reduce noise. The model's versatility in capturing additive relationships makes it applicable across various domains, where it provides insights into how combined factors drive observed outcomes.

Model Formulation

The sum-linear model is a classic linear factor model where each observable variable is represented as a sum of weighted factors plus an error term. Mathematically, it is expressed as:

$$\mathbf{X} = (X_1, \cdots, X_d) = \left(\sum_{i=1,\cdots,k} b_1^i Z_i, \cdots, \sum_{i=1,\cdots,k} b_d^i Z_i \right) + \epsilon \tag{2.3}$$

where $\mathbf{b}_i = (b_i^1, \cdots, b_i^d) \in [0, \infty)^d, i = 1, \cdots, k$ are k different factors, $Z_1, \cdots, Z_k$ are IID random variables with the same heavy-tailed distribution, for instance, with the standard Fréchet distribution being a most common choice, and ϵ is the $d-$dimensional random error. We also impose an assumption that

$$\sum_{i=1}^{k} b_j^i = 1 \text{ for all } j = 1, \cdots, d, \tag{2.4}$$

so that the marginal distributions of $\mathbf{X}$ have the same rate of tail decay, e.g., they are each standard Fréchet if Z_i's are assumed so.

Application and Challenges

The sum-linear model serves as a cornerstone of statistical modeling due to its straightforward structure and broad applicability across various fields. By expressing an observable outcome as a linear combination of contributing factors, this model provides a simple yet powerful means of understanding how multiple factors collectively influence a response variable. Developed initially to decompose complex systems into manageable components, the sum-linear model has proven invaluable in fields such as finance, econometrics, psychology, and machine learning, where it underpins many widely-used analytical methods. In finance, for example, asset pricing models like the Fama-French model use sum-linear models to capture how market risk, firm size, and value premiums impact asset returns. Similarly, in econometrics, sum-linear models enable policymakers to analyze macroeconomic indicators, allowing them to break down complex economic phenomena into interpretable factors that guide decision-making.

In the realm of data science and machine learning, the sum-linear model is foundational to techniques such as Principal Component Analysis (PCA) and factor analysis, which reduce data dimensionality and enhance interpretability. These techniques leverage the sum-linear approach to identify latent structures within data, removing noise and simplifying high-dimensional datasets. In psychology and social sciences, the model aids in survey analysis by explaining psychological constructs or behaviors as combinations of multiple latent factors, facilitating a deeper understanding of complex social behaviors and interactions.

Despite its strengths, the sum-linear model faces several challenges, particularly in applications involving complex, high-dimensional data. A fundamental limitation is its assumption of a strictly linear relationship between factors and the outcome, which can be overly simplistic when real-world relationships are non-linear. Additionally, multicollinearity among factors—where factors are highly correlated—can obscure individual factor contributions and lead to unreliable estimates, making interpretation challenging. Overfitting is another concern, especially in high-dimensional datasets where the model may capture noise rather than genuine patterns, compromising its predictive utility. Furthermore, the model's sensitivity to outliers can affect its robustness, as extreme values may disproportionately impact the fit.

Addressing these limitations requires careful consideration of model selection, validation procedures, and potentially robust estimation techniques.

Nonetheless, the sum-linear model remains an essential tool in statistical modeling, balancing interpretability and explanatory power. By providing a framework to assess additive effects, it continues to play a vital role in diverse applications, offering insights into how combined factors drive observed outcomes in complex datasets.

2.2 Spherical Clustering

To examine the behavior of extreme values, as previously discussed, it is essential to define the spectral measures. Spectral measures reveal the extremal dependence structure between variables, offering insights into how extreme values relate to one another. Clustering methods on the sphere can then be applied to capture and describe the distributional details of these dependencies.

Spherical clustering is a specialized form of clustering that organizes data points distributed on a spherical surface, such as in high-dimensional data where each point lies approximately on a hypersphere. This technique is especially useful when dealing with data that naturally has directional properties or can be represented using cosine similarity, like text data in natural language processing (NLP) or user preference vectors in recommender systems. Unlike traditional clustering methods, which rely on Euclidean distance, spherical clustering often uses cosine distance to measure similarity, making it suitable for cases where only the orientation of data vectors matters, not their magnitude. The spherical k-means algorithm is a prominent example in this domain, where centroids are updated to minimize cosine distance rather than Euclidean, leading to more meaningful clusters in high-dimensional, sparse spaces.

Additionally, spherical clustering is computationally efficient for large-scale text data and has been applied in topic modeling, document clustering, and semantic analysis. In these cases, text data represented in vector form using embeddings (e.g., Word2Vec or BERT) benefit from spherical clustering as it captures semantic relationships better than traditional methods. This method is also used in genomic data analysis, where genes can be clustered based on expression levels, which are often measured by similarity in direction rather than magnitude. However, spherical clustering has limitations, such as difficulty with non-spherical data distributions, as it assumes that clusters are evenly spread across the spherical space, potentially leading to poor performance when data clusters are dense or anisotropic. Recent advancements have focused on improving spherical clustering algorithms through techniques like spherical Gaussian mixtures and deep learning-based clustering, which can adaptively handle complex data distributions while retaining computational efficiency.

2.2.1 Relationship Between Model Estimation and Spherical Clustering

As defined previously, the spectral measure for the max-linear model and sum-linear model provide information on the extremal dependence structure between the variables in (2.1) and (2.3). From the structure of (2.1) and (2.3), it is known based on the single-big-jump principle of heavy-tailed distributions, the largest observations of $\mathbf{X}$ are due to a large observation of a Z_i and therefore the factors $\mathbf{b}_i$ determines the possible direction of extremal observations. In fact, it can be shown that the spectral measure S of the max-linear model or sum-linear model concentrates on the points $\mathbf{a}_i = \mathbf{b}_i / \|\mathbf{b}_i\|$ with corresponding probability $p_i = \|\mathbf{b}_i\| / (\sum_{l=1}^{k} \|\mathbf{b}_l\|),\ 1 \leq i \leq k$(Janßen and Wan, 2020a).

Figure 2.1 demonstrates simulations from a three-dimensional max-linear model. The factor $\mathbf{b}_i$'s are $\mathbf{b}_1 = (0.8, 0.1, 0.1)$, $\mathbf{b}_2 = (0.1, 0.8, 0.1)$, and $\mathbf{b}_3 = (0.1, 0.1, 0.8)$. There are 10,000 data points simulated based on a standard Frechèt distribution for each Z_i. The plot on the left shows all 10,000 data simulated, the plot in the center is 50% of the data set with the largest norm, and the one on the right is 10% of the data set with the largest norm. Apparently, the "extreme" points (points with the largest norm) are centered around the points $\mathbf{a}_1 = \mathbf{b}_1 / \|\mathbf{b}_1\| = (0.98, 0.12, 0.12)$, $\mathbf{a}_2 = \mathbf{b}_2 / \|\mathbf{b}_2\| = (0.12, 0.98, 0.12)$, and $\mathbf{a}_3 = \mathbf{b}_3 / \|\mathbf{b}_3\| = (0.12, 0.12, 0.98)$, labeled red in the plots. These points are roughly centered around $\mathbf{a}_i,\ i = 1, \cdots, 3$, with the same probability $p_i = 1/3$. Selecting more extremal portion of the data will make the spectral structure of the max-linear model more apparent.

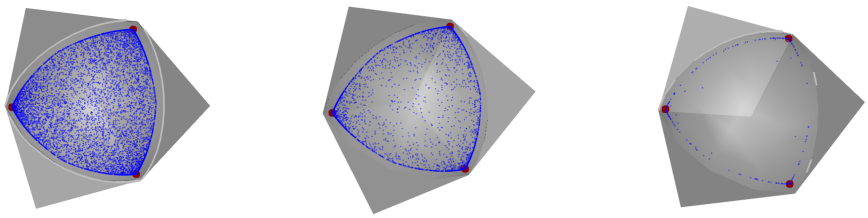


Figure 2.1: Simulated Max-linear Data (left: 100% data; middle: 50% data; right: 10% data)

Hence, the parameter estimation for max-linear models or sum-linear model can viewed as a clustering process. The points $\mathbf{a}_1, \cdots, \mathbf{a}_k \in \mathbb{S}_+^{d-1}$ are the centers where the spectral measure S concentrated with the corresponding probabilities $p_1, \cdots, p_k$. It was argued in Janßen and Wan, 2020a that the estimation

of $\mathbf{a}_i, i = 1, \cdots, k$ and $p_i, i = 1, \cdots, k$ obtained from a clustering procedure can be viewed as an alternative to the estimation of the factors $\mathbf{b}_i, i = 1, \cdots, k$. Note that, however, knowledge of $\mathbf{a}_i$ and $p_i, i = 1, \cdots, k$, (totally $dk - 1$ free parameters) is overdetermining $\mathbf{b}_i, i = 1, \cdots, k$, with the constraint (2.2) (totally $dk - k$ free parameters).

This thesis proposes two spherical clustering methods: spherical k-means clustering and spherical k-principal component clustering. Both methods offer effective approaches for estimating linear factor models. However, their applications may differ, particularly beyond model estimation. Notably, chapter 4 explores how these methods can be applied to detect concomitant extremes.

2.2.2 Spherical K-means Clustering

In high-dimensional data analysis, particularly in text mining and document clustering, traditional clustering algorithms such as k-means often face challenges due to the nature of the data. Specifically, text and other high-dimensional data are typically sparse, and the magnitude of feature vectors often has less significance compared to their directionality. To address this issue, a variant of the k-means algorithm, known as Spherical K-means Clustering, has been developed, which adapts to the characteristics of such data.

Spherical K-means clustering differs from the conventional K-means algorithm by using cosine similarity as the distance metric instead of Euclidean distance. Cosine similarity measures the cosine of the angle between two vectors, making it a more appropriate measure of similarity in high-dimensional spaces where the direction of the data points (rather than their magnitude) is of primary importance. This feature is particularly useful in domains like natural language processing (NLP), where documents or texts are represented as high-dimensional vectors, such as term-frequency inverse document frequency (TF-IDF) vectors or word embeddings.

As discussed in the previous sections, parameter estimation for linear factor models turns into a clustering problem. For such a problem, the spherical k-means algorithm, the spherical variant of the classical k-means algorithm, is a natural choice. More specifically, the spherical k-means clustering is a clustering technique that is used to group data points based on their similarity. Unlike the classical k-means clustering algorithm developed by Hartigan and Wong, 1979, which operates on Euclidean distances between data points, spherical k-means clustering works on the surface of a hypersphere (see Hornik et al., 2012b). The algorithm works on the surface of a hypersphere, rather than in an Euclidean space. The data points have been normalized to have unit distance to the origin, so that they lie on the surface of the hypersphere. The algorithm then proceeds

similarly to the traditional k-means algorithm, with some modifications to accommodate the spherical geometry.

Algorithm Description

In general, the algorithm begins by randomly selecting k initial cluster centers on the hypersphere, and then iteratively assigns each data point to its nearest cluster center based on their cosine similarity (see (2.5)). The centroid of each cluster is then updated based on the mean of the cosine dissimilarities between the data points assigned to the cluster and the cluster center. This process is repeated until convergence.

1. **Initialization**: The algorithm initializes by randomly selecting k initial cluster centroids $\{c_1, c_2, \ldots, c_k\}$. Each centroid is normalized to have unit length, i.e., $\|c_j\| = 1$ for all $j \in \{1, \ldots, k\}$, ensuring that the centroids lie on the unit hypersphere.

2. **Assignment Step**: For each data point x_i, the cosine dissimilarity between x_i and each centroid c_j is computed. The cosine dissimilarity is defined as:

$$\text{cosine\_dissimilarity}(x_i, c_j) := d(x_i, c_j)_\phi = 1 - \frac{x_i \cdot c_j}{\|x_i\|\|c_j\|} \tag{2.5}$$

   Since both x_i and c_j are normalized to unit length, this simplifies to:

$$\text{cosine\_dissimilarity}(x_i, c_j) = 1 - x_i \cdot c_j$$

   Each data point is then assigned to the cluster corresponding to the centroid with which it has the smallest cosine dissimilarity.

3. **Update Step**: After all data points have been assigned to clusters, the centroids are updated by computing the mean of all data points in each cluster. The new centroids are then normalized to lie on the unit hypersphere.

4. **Convergence**: The algorithm iterates between the assignment and update steps until the centroids stabilize or the change between consecutive iterations falls below a predefined threshold.

Here we provide a mathematical description of the objectives of the classical and the spherical k-means on the population level. The Let $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ be a dissimilarity function in $\mathbb{R}^d$. For a probability measure P on $\mathbb{R}^d$,

the averaged distance from any observation to the closest element of A can be represented as:

$$W(A, P) := \int_{\mathbb{R}^d} \min_{a \in A} d(x, a) P(dx) \in [0, \infty), \tag{2.6}$$

where $A = \{a_1, \cdots, a_k\}$, $a_i \in \mathbb{R}^d$ for $i = 1, \cdots, k$ and $k \in \mathbb{N}$. The k-means cluster center is a set A_k which minimizes $W(A, P)$ among all A. Spherical k-means, on the other hand, replaces $\mathbb{R}^d$ in (2.6) by $\mathbb{S}_+^{d-1}$ and P is now understood as a probability measure on $\mathbb{S}_+^{d-1}$, and uses angular dissimilarity in (2.5).

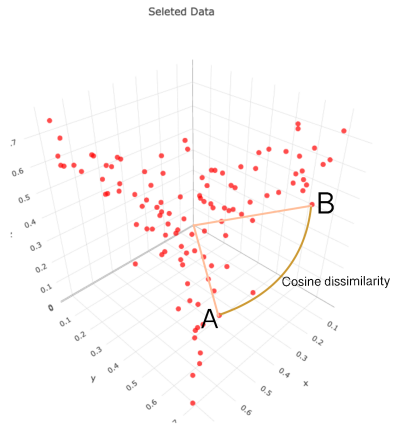


Figure 2.2: Cosine Dissimilarity

The optimal spherical k-means center set A_k is obtained similarly. In the practice of a (spherical) k-means algorithm, the population probability measure P is replaced by the empirical version say P_n, under which we denote the optimal center set as A_k^n. Proposition 3.3 in Janßen and Wan, 2020a provided a consistency result for the spherical k-means, showing the convergence of A_k^n to A_k as the sample size $n \to \infty$.

Advantages and Applications

One of the primary advantages of Spherical K-means is its effectiveness in handling high-dimensional, sparse data, which is a common feature of document-term matrices used in text mining. The normalization step and the use of cosine similarity ensure that the clustering process focuses on the directional alignment

of data points, making it more robust in cases where the magnitude of the vectors is less meaningful. This makes Spherical K-means particularly well-suited for applications such as *document clustering*, *topic modeling*, and *information retrieval*, where the goal is to group similar texts or documents based on their content.

For instance, in document clustering, a common approach is to represent each document as a vector of word frequencies. These vectors tend to be highly sparse, with many zero entries, and the differences in their magnitudes might not reflect the actual content similarity between the documents. By focusing on the angular distance between these vectors, Spherical K-means can more effectively group documents discussing similar topics.

Spherical k-means clustering has found applications in various domains where multi-dimensional data with angular relationships need to be clustered. There are a lot of notable applications of spherical k-means clustering. In natural language processing, text data is often represented as high-dimensional vectors, such as word embeddings or document-term matrices (See. e.g., Tunali et al., 2016). Spherical k-means clustering can be applied to cluster documents or words based on their semantic similarities, capturing the angular relationships between them. In computer vision, spherical k-means clustering has been used to cluster images based on visual features (See. e.g., Moriya et al., 2018). High-dimensional image descriptors, such as histograms or deep features, can be normalized and clustered on the hypersphere to capture the angular relationships between images. In genomics and bioinformatics, spherical k-means clustering has been applied to analyze high-dimensional gene expression data (See. e.g., Moussa and Măndoiu, 2018). By considering the angular relationships between gene expression profiles, it can identify clusters of genes with similar expression patterns, aiding in understanding biological processes and identifying potential biomarkers. The technique can be adapted to various domains where high-dimensional data with angular relationships need to be clustered, providing valuable insights and facilitating data-driven decision-making.

Limitations

Despite its advantages, Spherical K-means is not without its limitations. One significant assumption of the algorithm is that the data naturally lies on the surface of a hypersphere, which might not hold for all types of high-dimensional data. Additionally, like traditional K-means, Spherical K-means requires the number of clusters k to be specified *a priori*, which can be a challenge in unsupervised learning tasks where the true number of clusters is unknown. Furthermore, the

algorithm's performance is sensitive to the initial choice of centroids, which may result in suboptimal clustering in certain cases.

Summary

Spherical K-means is a powerful extension of the traditional K-means clustering algorithm, tailored for high-dimensional, sparse data. By leveraging cosine similarity and unit-length normalization, the algorithm provides a more meaningful clustering of data where the direction of the feature vectors is more important than their magnitude. This characteristic makes Spherical K-means particularly useful in text mining, document clustering, and other domains where data exhibits high dimensionality and sparsity. However, its assumptions and sensitivity to initialization must be carefully considered in practical applications.

2.2.3 Spherical K-principle Component Clustering

Spherical K-principal Component Clustering (SK-PC) is a clustering approach designed for high-dimensional data distributed on or near a spherical surface. It utilizes the theoretical foundation of principal component analysis (PCA) and leverages the properties of the first principal components, enhancing the clustering process. However, the primary focus of the method is on grouping similar data points rather than performing PCA. Unlike traditional clustering, which typically uses Euclidean distances, spherical clustering focuses on angular distances or cosine similarities, making it particularly suitable for data types where the direction of data vectors is more informative than their magnitude, such as in text data, genetic data, or other high-dimensional datasets with sparsity.

In Spherical K-Principal Component Clustering, principal component directions are computed to capture the primary modes of variation within clusters on the sphere. The clustering process is enhanced by projecting data points onto these principal components, and the first principle component can be seen as cluster centers or prototypes. This approach provides a more robust framework for clustering when the data's primary structure aligns with these principal directions. As a result, it enables more accurate modeling of the data's inherent structure, helping to mitigate noise and improving the interpretability of clustering results. Additionally, by emphasizing the orientation of data points rather than their absolute distances, this method can identify clusters that are more representative of the underlying data patterns, especially when applied to linear factor models or in detecting extreme events in high-dimensional spaces.

This clustering method is especially valuable in applications that require an understanding of extremal dependence, as it can help detect and interpret patterns among extreme values, such as identifying outliers or assessing risk in financial and environmental data.

2.2.4 Algorithm Description

The logistic structure of the k-pc algorithm is similar to spherical k-mean clustering. Instead of using means to find the optimal center in (2.6), the first principle direction calculated by a "covariance matrix" centered with respect to the origin is used (see Σ_i in Algorithm 4.2.2). Algrithm 4.2.2 gives a single iteration of spherical k-pc clustering.

1. **Input**: The algorithm initializes by randomly selecting k initial cluster centroids $\{c_1, c_2, \ldots, c_k\}$. Each centroid is normalized to have unit length, i.e., $\|c_j\| = 1$ for all $j \in \{1, \ldots, k\}$, ensuring that the centroids lie on the unit hypersphere.

2. **Step 1:** Compute the $n \times k$ matrix of dot products

$$M = (x_1, \cdots, x_k)(\theta_1 \cdots \theta_n)^T.$$

3. **Step 2:** Let v be the mean of row-wise maxima of M.

4. **Step 3:** For each row of M, find the index of the (first) maximal value and store them in g.

5. **Step 4:** For $i = 1$ to $i = k$, calculate $\Sigma_i = (1/n) \sum_{\mu=1}^{n} (\theta_\mu \theta_\mu^T 1_{\{g_\mu = i\}})$, and find the principal eigenvector.

6. **Output:** new centroids $\hat{x}_1, \cdots \hat{x}_k \in \mathbb{S}_+^{d-1}$ and the old value v.

Similar to spherical K-means clustering, the mathematical objective function remains the same as defined in Equation (2.6). However, instead of using cosine dissimilarity as defined in Equation (2.5), dissimilarity is measured using the first principal component.

Advantages and Applications

Spherical K-principal Component Clustering (SK-PC) represents an advanced method for clustering high-dimensional data by combining the principles of spherical clustering and principal component analysis (PCA). This technique is particularly advantageous for data where the direction of data vectors holds more significance than their magnitude, such as in sparse datasets commonly encountered in natural language processing (NLP) and genomic studies. By focusing on angular distances or cosine similarity, SK-PC emphasizes the relational orientation among data points, allowing for a meaningful reduction of dimensionality while preserving critical angular relationships. This approach also enhances interpretability by projecting data onto principal component directions that represent the primary modes of variation within each cluster, providing insights that may be obscured in traditional clustering methods relying on Euclidean distances.

The applications of SK-PC are broad, spanning fields that require clustering of complex, high-dimensional datasets with distinct directional patterns. In NLP, for instance, clustering word or document embeddings based on their principal components allows for more nuanced semantic groupings, improving tasks like topic modeling and document categorization. In finance and environmental sciences, SK-PC is particularly valuable for identifying extremal dependencies among variables, such as clustered patterns of extreme values that indicate correlated risk events. Additionally, in genomics, this method enables the clustering of gene expression profiles to reveal functional or regulatory relationships, which is crucial in identifying biomarkers and understanding biological networks. Thus, the integration of spherical clustering with PCA facilitates applications that require detailed analysis of data's directional structure, making it a versatile tool for extracting meaningful patterns in diverse high-dimensional contexts.

Limitations

Despite its advantages, SK-PC has certain limitations that must be considered. The method assumes that data points are distributed in a spherical or near-spherical manner, which may not hold across all types of datasets, potentially limiting clustering effectiveness when data structures deviate from this assumption. Moreover, SK-PC is sensitive to initial centroid selection, similar to other clustering algorithms, meaning results may vary depending on initializations. This sensitivity often requires multiple iterations to ensure robust clustering outcomes, increasing computational demand. Lastly, the computational com-

plexity of projecting data onto principal components and iteratively clustering can make SK-PC resource-intensive, particularly for very large datasets. Nevertheless, with careful tuning and application to suitable datasets, SK-PC remains a robust and insightful tool for understanding complex data structures.

Summary

Spherical K-Principal Component Clustering is a powerful tool for clustering high-dimensional, sparse, or directional data. By integrating principal component analysis with spherical clustering, it provides a more interpretable, directionally focused analysis that excels in applications such as NLP, financial risk modeling, and genomic studies. However, its effectiveness relies on assumptions about data distribution, and it can be computationally intensive. Despite these limitations, SK-PC offers significant value in applications where understanding extremal dependence and directionality is essential.

2.2.5 Comparison Between Spherical K Means Clustering and Spherical K Principle Component Clustering

Both Spherical K-means Clustering and Spherical K-Principal Component Clustering fall under the category of spherical clustering algorithms and can be summarized as follows.

The spherical clustering algorithms that have been considered so far are performed exclusively on the unit sphere $\mathbb{S}_+^{d-1}$ with respect to the 2-norm (Euclidean norm), that is, take $\|\cdot\|_{(s)}$ in (2.7) as $\|\cdot\|_2$.

$$\mathbb{S}_+^{d-1} = \{\mathbf{x} \in [0,\infty)^d : \ \|\mathbf{x}\|_{(s)} = 1\}, \tag{2.7}$$

We do not make this assumption for generality unless discussing specific examples. We equip $\mathbb{S}_+^{d-1}$ with the subspace topology inherited from $\mathbb{R}^d$. Next, we introduce a *dissimilarity measure* D that follows the assumption below.

Assumption 1 *Suppose* $D : \mathbb{S}_+^{d-1} \times \mathbb{S}_+^{d-1} \to [0,1]$ *is continuous, and satisfies the following properties: for* $\mathbf{w}_i \in \mathbb{S}_+^{d-1}$, $i \in 1,2$, *(i)* $D(\mathbf{w}_1, \mathbf{w}_2) = 0$ *if and only if* $\mathbf{w}_1 = \mathbf{w}_2$*; (ii)* $D(\mathbf{w}_1, \mathbf{w}_2) = D(\mathbf{w}_2, \mathbf{w}_1)$.

Remark 1 *Without loss of generality, we shall assume that* D *is properly normalized so that* D *is surjective over* $[0,1]$. *A nonnegative function* D *satisfying (i) and (ii) is often referred to as a* semimetric*, which lacks the triangular inequality*

axiom of a metric. With the assumptions imposed, we have $\mathbf{w}_n \to \mathbf{w}$ *on if and only if* $D(\mathbf{w}_n, \mathbf{w}) \to 0$ *as* $n \to \infty$*, and the* D*-neighborhoods*

$$B(\mathbf{w}, r) := \{\mathbf{u} \in\colon\ D(\mathbf{w}, \mathbf{u}) < r\},$$

$\mathbf{w} \in, r > 0$*, form a topological basis of ; see, e.g., Wilson, 1931,Galvin and Shore, 1984. Note that due to the compactness of and the continuity of* D*, the function*

$$D^\dagger(\mathbf{w}_1, \mathbf{w}_2) := \sup_{\mathbf{w}\in} |D(\mathbf{w}, \mathbf{w}_1) - D(\mathbf{w}, \mathbf{w}_2)| \tag{2.8}$$

is also a semimetric that is continuous on $\times$ *and maps surjectively to* $[0, 1]$*, which we refer to as the* dual *of* D*. Following from its definition, we have* $D^\dagger \geq D$*, and a triangular-like inequality holds:*

$$D(\mathbf{w}_1, \mathbf{w}_3) \leq D(\mathbf{w}_1, \mathbf{w}_2) + D^\dagger(\mathbf{w}_2, \mathbf{w}_3). \tag{2.9}$$

Some common dissimilarity measures are only semimetrics but not metrics. Below, we consider $\|\cdot\|_{(s)} = \|\cdot\|_2$ so that is the 2-norm sphere. The cosine dissimilarity adopted in the spherical k-means of Dhillon and Modha, 2001; Janßen and Wan, 2020b is given by

$$D_{\cos}(\mathbf{w}_1, \mathbf{w}_2) = 1 - \mathbf{w}_1^\top \mathbf{w}_2, \tag{2.10}$$

where $\mathbf{w}_1, \mathbf{w}_2 \in\subset \mathbb{R}^d$. The dissimilarity measure corresponding to the k-pc algorithm of Fomichov and Ivanovs, 2023 is given by

$$D_{\mathrm{pc}}(\mathbf{w}_1, \mathbf{w}_2) = 1 - \mathbf{w}_1^\top \mathbf{w}_2{}^2. \tag{2.11}$$

These two dissimilarity measures enjoy computational advantages, although neither of them is a metric. Note that since $\left|\mathbf{w}_1^\top \mathbf{w}_2{}^2 - \mathbf{w}_1^\top \mathbf{w}_3{}^2\right| \leq 2|\mathbf{w}_1^\top \mathbf{w}_2 - \mathbf{w}_1^\top \mathbf{w}_3| \leq 2\|\mathbf{w}_2 - \mathbf{w}_3\|_2$, $\mathbf{w}_i \in\subset \mathbb{R}^d$, one obtains a bound for the dual semimetric as $D^\dagger \mathbf{w}_2, \mathbf{w}_3 \leq c\|\mathbf{w}_2 - \mathbf{w}_3\|_2$ for $D = D_{\cos}$ or D_{pc}, with constant $c = 1$ or 2 respectively.

To simplify the mathematical description of clustering of sample data, it is convenient to use the notion of *multiset*. Recall that a multiset W on is a set that allows repetition of its elements, whose support, denoted as W, is a subset of in the usual sense that eliminates repetitions in W. For instance, with two distinct points $_1$ and $_2$ on , one can have $W = \{\mathbf{w}_1, \mathbf{w}_1, \mathbf{w}_2\}$ with $W = \{_{1,2}\}$. A multiset W can be characterized by the multiplicity function $m_W :\mapsto \{0, 1, 2, \ldots\}$, where $m_W(\mathbf{w})$ equals the number of repetitions of element $\mathbf{w} \in (m_W(\mathbf{w}) = 0$ if $\mathbf{w} \notin W)$. A subset of in the usual sense can

be understood as a multiset with the multiplicity taking value either 0 or 1, with the empty set corresponding to a multiplicity function that is identically 0. When the notation $\mathbf{w} \in W$ is used for a multiset W, it means that $\mathbf{w}$ is an element in W. For multisets W_1, W_2 with multiplicity functions m_1 and m_2 respectively, their union $W_1 \cup W_2$ is given by the multiset characterized by the multiplicity function $m_1 \vee m_2$, and their intersection $W_1 \cap W_2$ is given by the multiset characterized by $m_1 \wedge m_2$. The relation $W_1 \subset W_2$ is understood as $m_1 \leq m_2$. Furthermore, if W is a finite set, a summation $\sum_{\mathbf{w}\in W} f()$ for a suitable function f is understood as $\sum_{\mathbf{w}\in W} f() m_W(\mathbf{w})$. For example, the cardinality of W is defined as

$$|W| = \sum_{\in W} m_W(\mathbf{w}).$$

Also we write $D(\mathbf{w}, W) = \inf_{\mathbf{s}\in W} D(\mathbf{w}, \mathbf{s})$.

Now suppose W is a multiset on $\mathbb{S}_+^{d-1}$ with cardinality $|W| < \infty$. Suppose $k \in \mathbb{Z}_+$ and $k \leq |W|$. Let $A_k^* = \mathbf{a}_1^*, \ldots, \mathbf{a}_k^*$ be a multiset on with cardinality k, which satisfies

$$\sum_{\mathbf{w}\in W} D\mathbf{w}, A_k^* = \inf \left\{ \sum_{\mathbf{w}\in W} D(\mathbf{w}, A) : \ A \subset \mathbb{S}_+^{d-1}, \ |A| = k \right\}. \qquad (2.12)$$

The existence of A_k^* is guaranteed by the continuity of D and the compactness of $\mathbb{S}_+^{d-1}$, although it may not be unique. Notice that when $|W| \geq k$, the infimum in (2.12) must be achieved with a distinct set of $\mathbf{a}_i^*$'s. Below when multisets $C_1, \ldots, C_k$ with multiplicity functions $m_1, \ldots, m_k$ are said to form a partition of a multiset W with multiplicity function m, it means that $m = m_1 + \ldots + m_k$, and $m_i \neq 0$ for all $i \in \{1, \cdots, k\}$.

Definition 3 *A* k-clustering *of a multiset* W, $1 \leq k \leq |W|$, *with respect to the dissimilarity measure* D *refers to a pair* $(A_k^*, \mathfrak{C}_k)$. *Here* A_k^* *is as described above, and* $\mathfrak{C}_k = \{C_1, \ldots, C_k\}$ *is a partition of* W *into a collection of multisets* C_i*'s such that* $D\mathbf{w}, A_k^* = D\mathbf{w}, \mathbf{a}_i$ *for all* $\mathbf{w} \in C_i$, $i \in 1, \ldots, k$. *We refer to* A_k^* *as the* set of centers *and each* C_i *as a cluster.*

Remark 2 *A* k*-clustering of* W *always exists, although it may not be unique even when* A_k^* *is unique: there may be points in* W *with the same* D*-dissimilarity to multiple centers. On the other hand, it is always possible to ensure non-emptiness of each cluster* C_i *when* $k \leq |W|$.

With the choices $D = D_{\cos}$ and D_{pc} in (2.10) and (2.11), respectively, a k-clustering corresponds to the spherical k-means and k-pc clustering of Dhillon

and Modha, 2001 and Fomichov and Ivanovs, 2023, respectively. Solving a k-clustering problem can be computationally hard, and typically, the solution can only be approximated by a heuristic algorithm such as a Lloyd-type iterative algorithm as in Dhillon and Modha, 2001 and Fomichov and Ivanovs, 2023. In the theoretical analysis of this paper, we assume that a k-clustering can be found accurately. In addition, when W is later given by a random subsample W_n of the total sample $(\mathbf{X}_i)_{i=1,\dots,n}$, we assume that the elements in A_k^* and the labels $\mathbf{X}_i \in C_j$, $i \in \{1,\dots,n\}$, $j \in \{1,\dots,k\}$, are measurable.

In summary, spherical k-means clustering and spherical k-principal component clustering can be categorized as spherical clustering methods, as described in (2.6), distinguished by their choice of dissimilarity functions: cosine dissimilarity ((2.10)) for the spherical k-means method and principal component dissimilarity ((2.11)) for the spherical k-principal component method.

In Figure 2.3, an example comparing the spherical K-means clustering center (dark blue point) and the spherical k-principal component clustering center (red point) on the unit sphere illustrates the similarities and subtle differences between these two methods. The data are randomly generated on the unit sphere, and both methods are applied, with the spherical k-principal component method (SK-PC) using the first principal component to identify the center. Notably, the two centers are very close, indicating that the methods can be expected to exhibit similar behavior in many scenarios.

This similarity suggests that, in terms of practical applications, both methods may perform comparably when applied to tasks such as clustering or estimating linear factor models. However, the computational approaches differ: the spherical K-means method directly minimizes cosine dissimilarity, while the SK-PC method relies on the properties of the first principal component to determine the center. As a result, their computational efficiency and performance can vary slightly depending on the dataset and context.

In practical applications, both spherical K-means and spherical k-principal component clustering rely on approximation techniques for estimating the cluster centers. The spherical K-means method uses iterative updates to minimize cosine dissimilarity, providing an efficient but approximate solution to the clustering problem. Similarly, the spherical k-principal component method leverages the first principal component as a proxy for the direction of maximum variance, using this as an approximation for determining the cluster center.

While the proximity of the centers suggests similar performance in many cases, the methods' underlying approximations can lead to slight differences in computational efficiency and clustering results. The spherical K-means method

focuses solely on minimizing angular dissimilarity, while the SK-PC method may introduce variations depending on the data structure. These differences highlight the trade-offs inherent in both approaches, as neither provides an exact solution but instead offers computationally practical approximations tailored to different contexts.

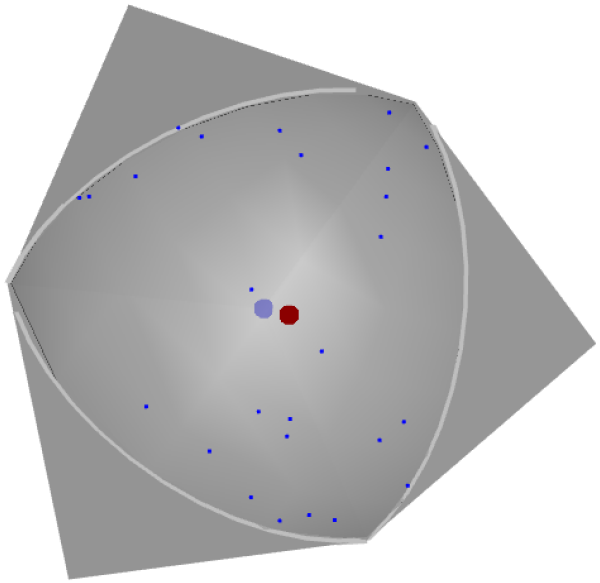


Figure 2.3: Spherical K-means Center (Dark blue) and Spherical K-principle Component Center (Red)

Chapter 3

On estimation and order selection for multivariate extremes via clustering

3.1 Model Estimation for Linear Factor Models

Back to the linear factor models, as observed by Einmahl et al., 2012 and Janßen and Wan, 2020b, one may relate a spherical k-clustering algorithm to the estimation of certain factor-like models that are often considered in the analysis of multivariate extremes. We summarize the linear factor models in the following content.

Suppose $B = (b_{ij})_{i=1,\dots,d,j=1,\dots,k} = \mathbf{b}_1, \dots, \mathbf{b}_k$, where $\mathbf{b}_j = (b_{1j}, \dots, b_{dj})^\top$, $j \in \{1, \dots, k\}$, are k distinct d-dimensional vectors, $b_{ij} \geq 0$, and that each column and row vector of B is nonzero (otherwise, the dimension d or the factor order k can be reduced).

Assume that $\mathbf{Z} = (Z_1, \dots, Z_k)^\top$ is a vector of i.i.d. positive continuous random variables satisfying $\Pr(Z_1 > z) \sim z^{-\alpha}$ as $z \to \infty, \alpha \in (0, \infty)$. Then the sum-linear model is given as

$$\mathbf{X} = X_1, \dots, X_d{}^\top = \sum_{j=1}^{k} b_{1j} Z_j, \dots, \sum_{j=1}^{k} b_{dj} Z_j^\top = B\mathbf{Z}. \tag{3.1}$$

On the other hand, we also have the max-linear model as

$$\mathbf{X} = X_1, \ldots, X_d{}^\top = \bigvee_{j=1}^{k} b_{1j} Z_j, \ldots, \bigvee_{j=1}^{k} b_{dj} Z_j{}^\top = B \odot \mathbf{Z}, \qquad (3.2)$$

where $\odot$ is interpreted as the matrix product with the sum operation replaced by the maximum operation. Note that due to the exchangeability of $Z_1, \ldots, Z_k$, either model is identifiable only up to a permutation of the vectors $\mathbf{b}_j$, $j \in \{1, \ldots, k\}$, i.e. the distribution of $\mathbf{X}$ is unchanged if B is replaced by $B_\pi := \mathbf{b}_{\pi(1)}, \ldots, \mathbf{b}_{\pi(k)}$ for any permutation $\pi : \{1, \ldots, k\} \mapsto \{1, \ldots, k\}$. The models of types (3.1) and (3.2) have recently attracted considerable interest in connection with causal structural equations for extremes; see, e.g., Gissibl and Klüppelberg, 2018; Gnecco et al., 2021.

It is known that both models have a discrete spectral measure as in (1) and the spectral measure of linear factor models has the form:

$$H = \sum_{i=1}^{k} p_i \delta_{a_i}, \qquad (3.3)$$

where a_i's are distinct points on $\mathbb{S}^{d-1}$, and $p_i > 0, p_1 + \ldots + p_k = 1$. Here

$$p_j = \frac{\|\mathbf{b}_j\|}{\sum_{\ell=1}^{k} \|\mathbf{b}_\ell\|}, \quad \mathbf{a}_j = \frac{\mathbf{b}_j}{\|\mathbf{b}_j\|}, \quad j \in 1, \ldots, k. \qquad (3.4)$$

This can be derived based on the well-known "single large jump" heuristic: when $\|\mathbf{X}\|_{(r)}$ is large, it is only due to a single large Z_j with overwhelming probability. See, e.g., Medina et al., 2021 and Einmahl et al., 2012; we mention that these works usually assume the same norm $\|\cdot\|_{(r)} = \|\cdot\|_{(s)}$ and $\alpha = 1$, although an extension is straightforward. In addition, the marginal standardization condition imposes the following restriction on B:

$$\sum_{j=1}^{k} b_{ij} = 1, \quad i \in 1, \ldots, d. \qquad (3.5)$$

We also mention that one may relax the models (3.1) and (3.2) by adding a noise term, e.g., $\mathbf{X} = B\mathbf{Z} + \boldsymbol{\varepsilon}$ or $\mathbf{X} = (B \odot \mathbf{Z}) \vee \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)^\top$ is a vector of i.i.d. positive noise terms, and the maximum $\vee$ is performed coordinate-wise. As long as each ε_i has a tail lighter than that of Z_j, the conclusions made above still hold (see, e.g., Einmahl et al., 2012). The discussion also applies to the transformed-linear model of Cooley and Thibaud, 2019. Finally,

we mention that in the context of multivariate extremes, one typically only considers fitting these models to an extremal subsample instead of the whole sample.

Due to the discrete nature of the spectral measure, the likelihood functions of these models are inaccessible (see, e.g., Einmahl et al., 2012, 2018; Yuen and Stoev, 2014). Even without taking a perspective of extremes, the max-linear model does not admit a smooth density.

Suppose from now on the order k is assumed to be known. Another noteworthy issue deserving discussion is whether we can translate the estimation of the spectral measure through a k-clustering algorithm into an estimation of the coefficient matrix $B = \mathbf{b}_1, \ldots, \mathbf{b}_k$ in (3.1) or (3.2). Note that the constraint (3.5) also needs to be taken into account. Combining (3.4) and (3.5), to solve the kd coefficients in B from p_j's and $\mathbf{a}_j$'s, we have totally $kd + d - 1$ free equations ($k - 1$ from the equations for p_j's, $(d-1)k$ from the equations for $\mathbf{a}_j$'s and d from (3.5)). When p_j's and $\mathbf{a}_j$'s are estimated via k-clustering, the over-determined system may not admit a solution, although this over-determined relation holds asymptotically.

In the following, we describe a simple method to convert spectral estimation to an estimation of B that satisfies the constraint (3.5). Observe that the exponent measure Λ for the models (3.1) and (3.2) concentrates on the rays $\{t\mathbf{b}_j : t > 0\}$, $j \in 1, \ldots, k$. Hence a spectral mass point $\mathbf{a}_j = \mathbf{b}_j/\|\mathbf{b}_j\|$ on the $\|\cdot\|$-norm sphere corresponds to a spectral mass point $\mathbf{b}_j/\|\mathbf{b}_j\|_\alpha = \mathbf{a}_j/\|\mathbf{a}_j\|$ on the l_2-norm sphere, $j \in 1, \ldots, k$. The advantage of considering the l_2-norm sphere is that

$$\sum_{j=1}^{k} \|\mathbf{b}_j\| = \sum_{i=1}^{d} \sum_{j=1}^{k} b_{ij} = d$$

due to relation (3.5). Therefore, under the choice $\|\cdot\|_{(r)} = \|\cdot\|$ in (3.4), we have $p_j d = \|\mathbf{b}_j\|$, and hence

$$\mathbf{b}_j = (p_j d)^{1/2} \frac{\mathbf{a}_j}{\|\mathbf{a}_j\|}, \quad j \in 1, \ldots, k. \tag{3.6}$$

Note that one can plug in estimated $\mathbf{a}_j$ and p_j via k-clustering on the α-norm sphere into (3.6), obtaining, say, $\mathbf{b}_j$, $j \in \{1, \ldots, k\}$. However, the condition (3.5) may not be satisfied. We propose the following simple correction: first, form the preliminary estimated coefficient matrix $B := \mathbf{b}_1, \ldots, \mathbf{b}_k =: \mathbf{r}_1, \ldots, {\mathbf{r}_d}^\top$, where $\mathbf{r}_i^\top$, $i \in 1, \ldots, d$, are row vectors of B. Then we obtain the final estimate $B = \mathbf{b}_1, \ldots, \mathbf{b}_k$ of B through replacing each row $\mathbf{r}_i$ by $\mathbf{r}_i/\|\mathbf{r}_i\|_\alpha$, which ensures (3.5). It follows from a continuous mapping argument that the

thus obtained estimate of B is consistent (up to a permutation of $\mathbf{b}_i$'s). **Corollary 2.6.** from Deng et al., 2024, we provided the proof of consistency.

3.2 Order Selection for Linear Factor Models

Order selection is critical in factor models because it determines the number of factors that best represent the underlying structure of the data, directly impacting model accuracy, interpretability, and computational efficiency. Choosing the optimal number of factors ensures that the model captures essential patterns without overfitting, which can introduce noise and reduce predictive power. Proper order selection enhances the model's ability to explain variability in the data effectively, aiding in robust factor analysis, dimensionality reduction, and clustering applications. Techniques like penalized criteria (e.g., penalized silhouette methods in clustering) have become essential for achieving reliable order selection, particularly in high-dimensional and complex datasets.

3.3 Heuristic Approaches

Clustering is often an exploratory data analysis technique, and the true underlying structure of the data may be unknown. Without a ground truth, it becomes subjective to determine the optimal number of clusters. Many methods are proposed; however, it is important to be aware that clustering is not a definitive solution, and the results should be interpreted and validated carefully. In this proposal, we shall first discuss two heuristic approaches: the elbow method and silhouette analysis, to compare with the cross-validation method applied in next section with solid theoretical background.

Elbow Method

The elbow method (Bholowalia and Kumar, 2014) is a graphical technique that plots the number of clusters against a clustering evaluation metric (e.g., the minimized mean distance) and looks for a point where the decrease in the metric slows down significantly. This point is referred to as the "elbow" and can indicate the optimal number of clusters.

3.3.1 Gap Statistics

The Gap Statistic Tibshirani et al., 2001 evaluates clustering performance by comparing the total within-cluster variation of the observed data to that of

reference datasets, aiming to maximize the difference, or "gap," between them. While this method offers an objective criterion, it often necessitates visual plot interpretation, making the selection process semi-quantitative. However, in the context of factor models, the complexity of expressing the likelihood function and the need for a suitable reference distribution make applying the Gap Statistic challenging. Consequently, obtaining reliable gap statistics for the linear factor models discussed in this chapter is not straightforward.

Silhouette Analysis

Silhouette analysis (Rousseeuw, 1987) measures the compactness and separation of clusters. It computes a silhouette coefficient for each data point, which indicates how well it belongs to its assigned cluster compared to neighboring clusters. The average silhouette score across different numbers of clusters can help identify the optimal number with the highest overall cohesion and separation. The silhouette analysis provides a quantitative measure of the clustering quality for different numbers of clusters. Higher silhouette scores indicate better-defined clusters, while lower scores suggest that samples may be assigned to incorrect or ambiguous clusters.

The silhouette method calculates two main values for each point: the average distance to points in the same cluster $a(i)$ and the average distance to points in the nearest neighboring cluster $b(i)$. The silhouette scores $s(i)$ for each point is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3.7}$$

This metric is computationally intensive, especially for large datasets, as it requires distance calculations between each point and all other points in both the same and neighboring clusters.

To address the computational challenges of the silhouette method, researchers have developed the simplified silhouette method. This adaptation seeks to reduce the computational load while retaining the essential characteristics of the original silhouette metric. The simplified approach approximates the distance calculations, focusing on a representative subset of points or using efficient data structures to estimate distances. While this method may sacrifice some precision, it enables faster evaluation of clustering quality in large-scale datasets, making it particularly useful in big data scenarios where computational efficiency is crucial.

Both the silhouette and simplified silhouette methods provide valuable insights into the clustering structure, offering a balance between accuracy and

computational feasibility. Their use in a thesis on clustering highlights the importance of evaluating clustering performance, especially when dealing with complex or high-dimensional datasets. The choice between the classic silhouette method and its simplified counterpart often depends on the dataset's size and the computational resources available, as well as the desired balance between evaluation accuracy and processing time.

3.3.2 Silhouette Method with Penalty

The method we proposed are based on the silhouette method. We first address the simplified silhouette method in our linear factor model and the spherical clustering context.

Let $A_k^* = a_1^*, \ldots, a_k^*, \mathfrak{C}_k = C_1, \ldots, C_k$ be a k-clustering of W with respect to a dissimilarity measure D as in Definition 3.

Define for $w \in W$ that

$$a(w) = D(w, A_k^*), \quad \text{and} \quad b(w) = \bigvee_{i=1}^{k} D(w, A_k^* \setminus a_i^*),$$

which are respectively the dissimilarities of w to the closest center (i.e., the center of the cluster it belongs to) and to the second closest center. When $k = 1$. we understand $b(w) = 1$.

The (simplified) average silhouette width (ASW) Hruschka et al., 2004 of this k-clustering is then defined as

$$\bar{S} = \bar{S}\left(W; A_k^*\right) = \frac{1}{|W|} \sum_{w \in W} \frac{b(w) - a(w)}{b(w)} = 1 - \frac{1}{|W|} \sum_{w \in W} \frac{a(w)}{b(w)}. \quad (3.8)$$

A well-clustered dataset is expected to have small $a(w)$ values relative to $b(w)$ across the majority of w points.

Hence, one often uses $\bar{S}$ to guide the selection of the number of clusters, that is, to choose k which maximizes $\bar{S}$. However, when experimenting applying the ASW to multivariate extremes with a discrete spectral measure as described previously, the performance is unsatisfactory: it tends to respond insensitively when the number of clusters exceeds the true k, i.e., the number of atoms of the spectral measure. In particular, we observe two behaviors of ASW that lead to the issue: 1) it tends to treat a tiny fraction of isolated points as a cluster; 2) it sometimes splits a single cluster center into multiple centers that are close to each other.

To improve the method and aviod the issues addressed, the penalized ASW is given by

$$s_t = \bar{s} - p_t = \left(\frac{\min_i |C_i|}{|W|/k}\right)^t \left(\min_{1\le i<j\le k} D\left(a_i^*, a_j^*\right)\right)^t - \frac{1}{|W|}\sum_{\mathbf{w}\in W}\frac{a(\mathbf{w})}{b(\mathbf{w})}. \tag{3.9}$$

choose k which maximizes s_t.

Theorem 3.1 in Deng et al., 2024 implies that as long as the tuning parameter is in an appropriate range, with probability tending to 1 as $n \to \infty$, the true order $m = k$ uniquely maximizes the penalized ASW.

In practice, we suggest plotting the penalized ASW S_t as a function of $m = 1, 2, \ldots$, for a range of small t values. The idea is to start with t near 0, gradually increase it, and see how the penalized ASW curve responds. If some obvious spurious clusters (i.e., those with tiny size or centers that are too close together) are present, the curve tends to respond sensitively and bends at the appropriate order. We then identify the turning point m as the choice of the order k.

As a quick illustration, we follow a simulation setup of $(d = 6, k = 6)$ below to simulate a max-linear factor model. Penalized Average Silhouette Width (ASW) S_t (vertical axis) for spherical k-means clustering is plotted as a function of test order m (horizontal axis). The true discrete spectral measure are given as the following:

- $(a_1, p_1) = ((0.29, 0.21, 0.50, 0.45, 0.43, 0.49)^\top, 0.22)$,
- $(a_2, p_2) = ((0.74, 0.00, 0.59, 0.00, 0.32, 0.00)^\top, 0.10)$,
- $(a_3, p_3) = ((0.00, 0.27, 0.00, 0.47, 0.00, 0.84)^\top, 0.13)$,
- $(a_4, p_4) = ((0.33, 0.70, 0.63, 0.00, 0.00, 0.00)^\top, 0.14)$,
- $(a_5, p_5) = ((0.00, 0.00, 0.00, 0.81, 0.47, 0.34)^\top, 0.09)$,
- $(a_6, p_6) = ((0.48, 0.49, 0.25, 0.33, 0.53, 0.29)^\top, 0.32)$.

See Figure 3.1 . Increasing t to a very large value is not informative and is not recommended in practice. On the other hand, it would be desirable to develop a data-driven method for choosing t, which we leave for a future work to explore.

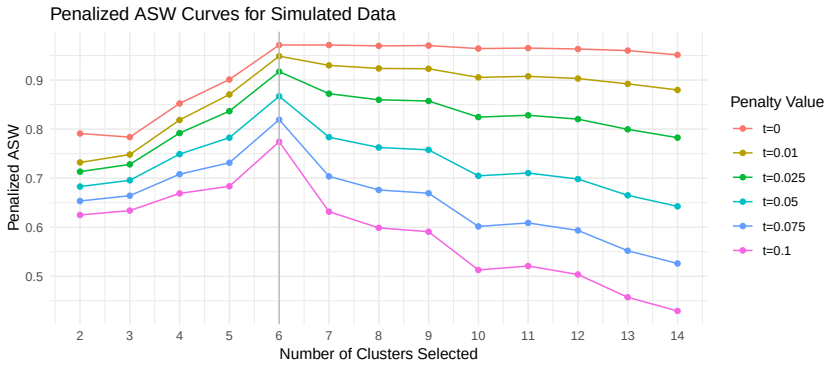


Figure 3.1: A simulation instance taken from $d = 6, k = 6$ setup.

3.4 Simulation and real data studies

3.4.1 Simulation Studies

In this section, we present some simulation studies to illustrate the performance of the penalized ASW method introduced. We follow the setup in Janßen and Wan, 2020b, Section 4 to simulate the max-linear factor model (3.2) with randomly generated coefficient matrix B. In particular, we let the factors Z_j's each follow a standard Fréchet ($\alpha = 1$) distribution. We consider 4 different combinations of dimensionality d and true order k. Under each (d, k) combination, we describe in the list below the way the coefficient vector $\mathbf{b}_j$'s are generated. Note that due to the standardization (3.5), only $\mathbf{b}_1, \ldots, \mathbf{b}_{k-1}$ need to be specified. Let U_i's stand for i.i.d. uniform random variables on $[0, 1]$.

- $d = 4, k = 2$: $\mathbf{b}_1 = (U_1, U_2, U_3, U_4)^\top/2$.
- $d = 4, k = 6$: $\mathbf{b}_1 = (U_1, U_2, U_3, U_4)^\top/3$, $\mathbf{b}_2 = (U_5, 0, U_6, 0)^\top/3$, $\mathbf{b}_3 = (0, U_7, 0, U_8)^\top/3$, $\mathbf{b}_4 = (U_9, U_{10}, 0, 0)^\top/3$, $\mathbf{b}_5 = (0, 0, U_{11}, U_{12})^\top/3$.
- $d = 6, k = 6$: $\mathbf{b}_1 = (U_1, \cdots, U_6)^\top/3$, $\mathbf{b}_2 = (U_7, 0, U_8, 0, U_9, 0)^\top/3$, $\mathbf{b}_3 = (0, U_{10}, 0, U_{11}, 0, U_{12})^\top/3$, $\mathbf{b}_4 = (U_{13}, U_{14}, U_{15}, 0, 0, 0)^\top/3$, $\mathbf{b}_5 = (0, 0, 0, U_{13}, U_{14}, U_{15})^\top/3$.
- $d = 10, k = 6$: First 5 factors are $\mathbf{b}_1 = (U_1, \cdots, U_{10})^\top/2$, $\mathbf{b}_2 = (U_{11}, U_{12}, 0, \cdots, 0)^\top/2$, $\mathbf{b}_3 = (0, 0, U_{13}, U_{14}, 0, \cdots, 0)^\top/2$, $\mathbf{b}_4 = (0, 0, 0, 0, U_{15}, U_{16}, 0, 0, 0, 0)^\top/2$, $\mathbf{b}_5 = (0, \cdots, 0, U_{17}, U_{18}, U_{19}, U_{20})^\top/2$.

For each of the 4 simulation setups described above, we randomly generate 100 models (i.e, 100 coefficient B matrices). From each of these generated models, we simulate a dataset of size 1000, extract a subsample of size 100 with the largest 2-norms, and project the subsample on the 2-norm sphere, namely, we work with $\|\cdot\|_{(r)} = \|\cdot\|_{(s)} = \|\cdot\|_2$. Subsequently, a spherical clustering algorithm (spherical k-means or k-pc) and the computation of the penalized ASW score is carried out on this projected subsample. Throughout the paper, for the spherical k-means algorithm, we use the implementation in the R package `skmeans` Hornik et al., 2012a, and for the k-pc algorithm, we use the R implementation provided in the supplementary material of Fomichov and Ivanovs, 2023.

In Figures 3.2 $\sim$ 3.5, we demonstrate the simulation results through some graphical representations. Specifically, each colored matrix plot is associated with a (d, k) setup as described above. In each plot, a column corresponds to a simulated dataset, and there are 100 columns. The upper half of the plot corresponds to spherical k-means and the lower half corresponds to k-pc. Within each of these halves, a row corresponds to a t penalty parameter specification. The color of a cell in the matrix signifies the order m chosen by maximizing the penalized ASW. We use a white color to indicate a coincidence of m with the true order k, with a deeper shade of red indicating that the greater m falls below the true k, and a deeper shade of blue indicating the greater it exceeds the true k. The bar graph to the right of the matrix indicates the success rate of order identification (that is, $m = k$) in all 100 instances.

In all these simulation setups, we can observe a tendency for the non-penalized ($t = 0$) ASW to overestimate (sometimes greatly) the order. As the penalty parameter t is tuned up from 0, we observe a significant bias correction effect, and the order identification success rate is noticeably improved over a range of $t > 0$. Note that this success rate is calculated with respect to the same t for different simulated data sets. We expect the success rate to improve if t is adaptively tuned for each dataset following the visual method described. It is also worth mentioning that the order identification based on k-pc tends to be more accurate than that based on k-means in most of these simulations.

3.4.2 Real data demonstrations

In this section, we use real data examples to demonstrate order selection through penalized ASW as introduced, as well as conversion of clustering-based spectral estimation to a factor coefficient matrix as mentioned. We present only the analysis based on the spherical k-pc algorithm, that is, the dissimilarity measure D is as in (2.11). The reason for doing so is two-fold. Firstly, the simulation study seems to suggest a better empirical performance for order selection based

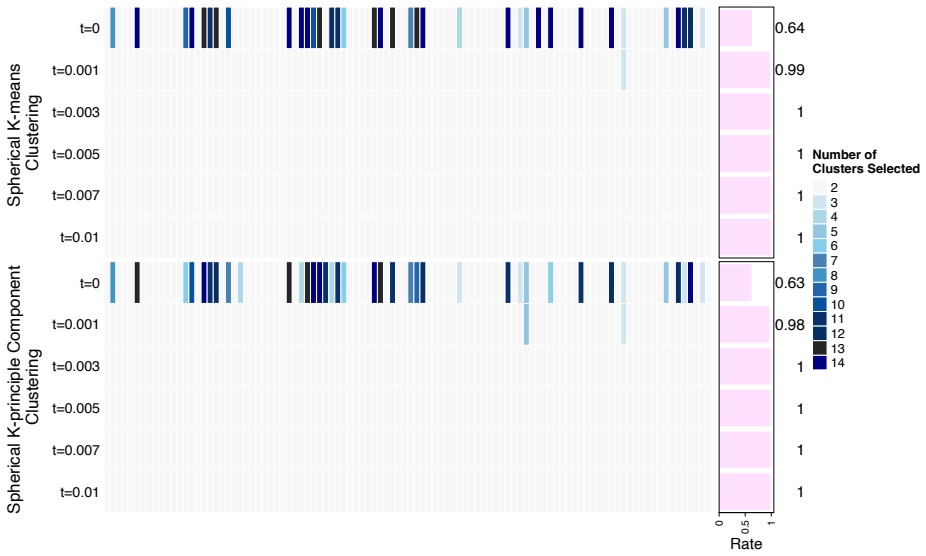


Figure 3.2: Simulation result visualization for the setup $d = 4, k = 2$.

on the k-pc algorithm. Secondly, as pointed out in Fomichov and Ivanovs, 2023, the k-pc algorithm is more suitable for the detection of groups of concomitant extremes, namely, subsets of variables that tend to be simultaneously large. The second property facilitates the comparison of the order k selected with some "ground truth" from the background information of the datasets.

In each of these studies, suppose that the observed data is $(\mathbf{x}_i) = (\mathbf{x}_i = x_{i1}, \ldots, x_{id}{}^\top \in [0, \infty)^d,\ i \in \{1, \ldots, n\})$. We follow a conventional approach to marginally standardize a dataset, so that the assumption with $\alpha = 2$ is roughly met. In particular, setting $\hat{F}_j(x) = n^{-1} \sum_{i=1}^{n} x_{ij} < x$ (under this choice of empirical CDF we ensure $\hat{F}_j(x_{ij})) < 1$), $j \in \{1, \ldots, d\}$, the transformed data is given by $(\mathbf{x}_i) = (\mathbf{x}_i = x_{i1}, \ldots, x_{id}{}^\top \in [0, \infty)^d,\ i \in \{1, \ldots, n\})$, where $x_{ij} := -\log \hat{F}_j(x_{ij})^{-1/2}$; if $\hat{F}_j$ were the true CDF for the data in dimension j, then x_{ij} would follow a standard 2-Fréchet distribution. Next, to prepare for the clustering of multivariate extremes, as in the simulation study, we select the extremal subsample of $(\mathbf{x}_i)$ with 10% largest 2-norms and project the subsample onto the 2-norm sphere, namely, we work with $\|\cdot\|_{(r)} = \|\cdot\|_{(s)} = \|\cdot\|_2$.

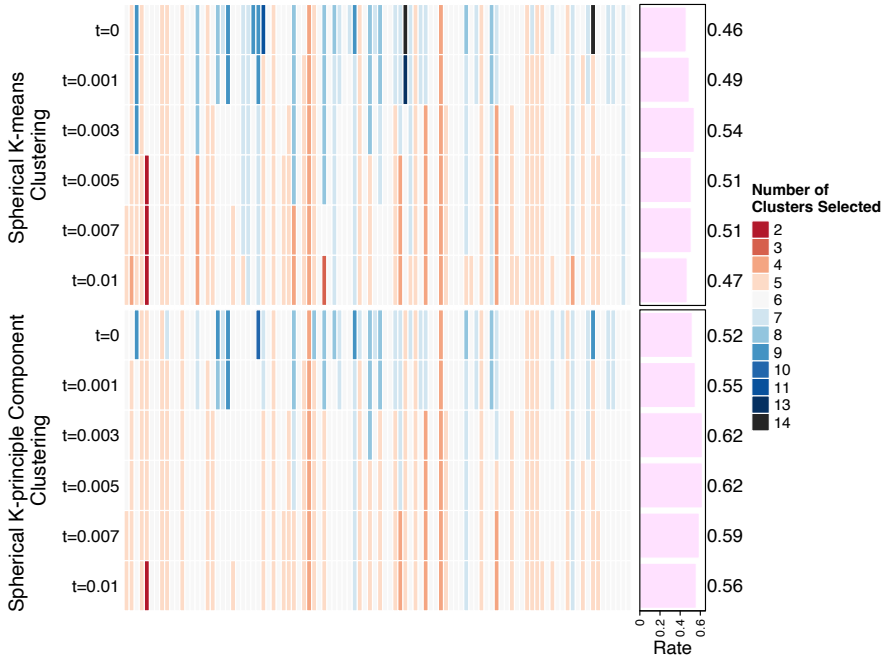


Figure 3.3: Simulation result visualization for the setup $d = 4, k = 6$.

Air Pollution Data

The air pollution dataset is found in the R package **texmex**, orginated from an online supplementary material of Heffernan and Tawn, 2004. It concerns air quality recordings in Leeds, U.K., specifically in the city center. The data span from 1994 to 1998, divided into summer and winter sets. The summer dataset comprises 578 observations, covering the months from April to July inclusively, while the winter dataset consists of 532 observations, encompassing the months from November to February inclusively. Each observation records the daily maximum values of five pollutants: Ozone, NO_2, NO, SO_2 and PM10. These datasets were also used in Janßen and Wan, 2020b to demonstrate the application of the spherical k-means clustering method to multivariate extremes.

In Figures 3.6, 3.7, 3.8 and 3.9, the penalized ASW is plotted against the number of clusters, where different curves correspond to different values of the tuning parameter t. With the visual method described, we can identify orders as 5 for the summer data and 3 for the winter data respectively.

These orders are similar to the choices 5 for the summer data and 4 for the winter data made in Janßen and Wan, 2020b under the guidance of certain elbow plots (see Janßen and Wan, 2020b, Figure 1). The authors did not provide

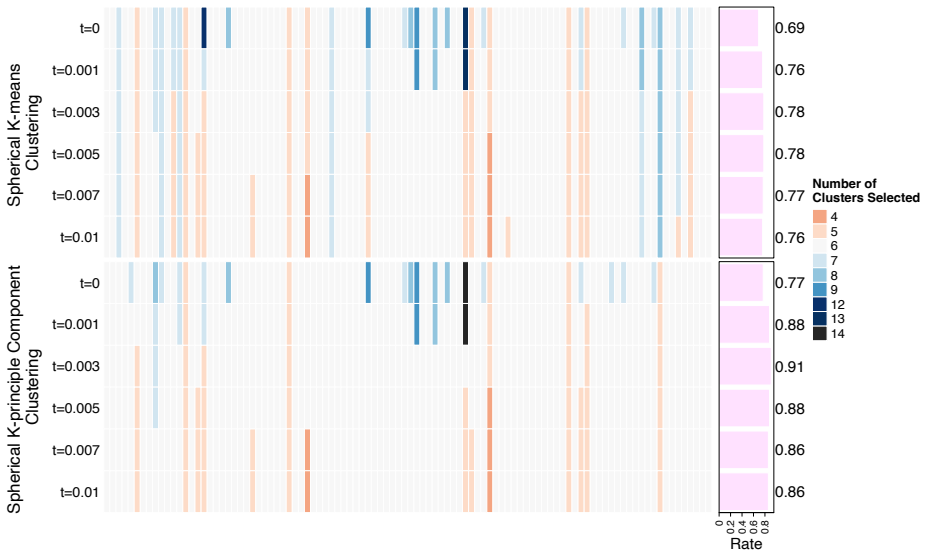


Figure 3.4: Simulation result visualization for the setup $d = 6, k = 6$.

a precise explanation of their choices. From the elblow plot in Janßen and Wan, 2020b, Figure 1, it seems that $k = 3$ for the winter data is also plausible. Recall also that here we use the spherical k-pc algorithm of Fomichov and Ivanovs, 2023 while Janßen and Wan, 2020b used the spherical k-means.

Furthermore, Figures 3.10 and 3.11 include visualizations of cluster centers computed based on the k-pc algorithm of Fomichov and Ivanovs, 2023 for the two datasets when we choose the numbers of clusters as above, respectively. Each row in either of the plots corresponds to the coordinate vector of a cluster center: a deeper shade of color indicates a higher value of the squared coordinate. Note that since we work with the 2-norm sphere, the squared coordinates for each cluster center sum up to 1, forming a probability distribution row-wise. For the summer data in Figure 3.7, whose order has been chosen as 5, the cluster centers concentrate sharply near coordinate directions, which to an extent indicates an asymptotic (or say extremal) independence (see, e.g., Beirlant et al., 2006, chapter8) of the pollutants.

In contrast, for the winter data in Figure 3.9, whose order has been chosen as 3, a cluster center indicates a group of concomitant extremes consisting of NO, NO2 and PM10. The asymptotic dependence between these 3 variables has been observed in Heffernan and Tawn, 2004. This serves as a support for our order choice which has placed these 3 variables in the same concomitant group.

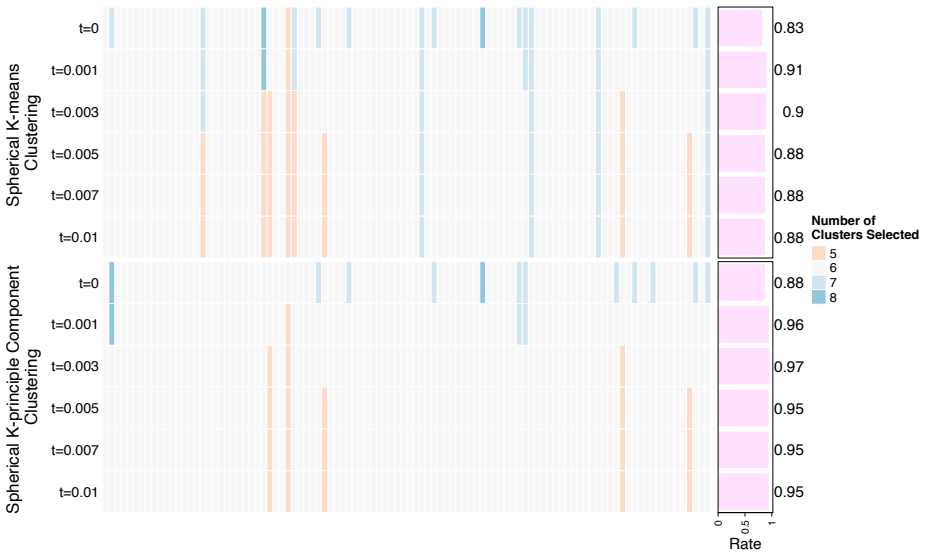


Figure 3.5: Simulation result visualization for the setup $d = 10, k = 6$.

Following the method introduced with $\|\cdot\|_{(s)} = \|\cdot\|_{(r)} = \|\cdot\|_2$ and $\alpha = 2$, we compute the factor coefficient matrix B for the two datasets; see Tables 3.1 and 3.2.

Table 3.1: Estimated $B^\top$ for Summer Pollution Data

Factor	O_3	NO_2	NO	SO_2	PM_{10}
1	0.88	0.22	0.10	0.20	0.24
2	0.20	0.33	0.20	0.90	0.32
3	0.35	0.79	0.30	0.21	0.32
4	0.15	0.16	0.16	0.19	0.80
5	0.21	0.44	0.91	0.25	0.31

River Discharge Data

The river discharge data concerns the daily discharge rate of rivers in North America sourced from the Global Runoff Data Centre German Federal Institute of Hydrology, n.d. The dataset comprises 16,386 daily records of discharge values from 13 stations spanning the period from December 1, 1976, to October 11, 2021. These 13 stations, shown in Table 3.3 and Figure 3.12, are positioned along 5 rivers in America: Willamette River, Mississippi River, Williamson River, Hudson River, and Broad River.

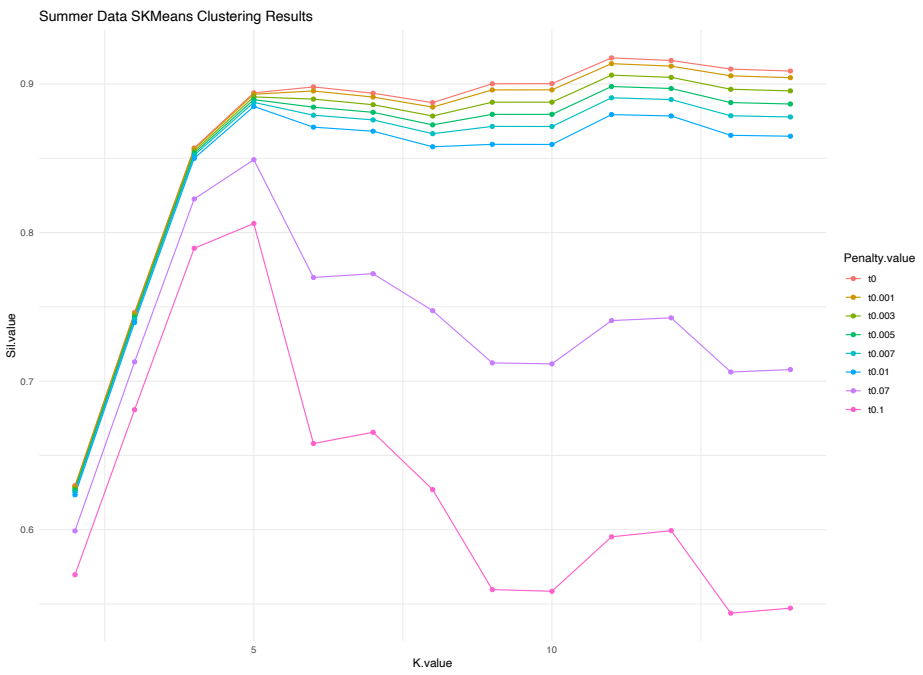


Figure 3.6: Air Pollution Example Summer-Spherical K Means Clustering

Table 3.2: Estimated $B^\top$ for Winter Pollution Data

Factor	O_3	NO_2	NO	SO_2	PM_{10}
1	0.19	0.98	0.99	0.44	0.98
2	0.07	0.13	0.12	0.89	0.14
3	0.98	0.12	0.10	0.07	0.15

As in the previous example, Figure **??** and 3.14 presents the penalized ASW curves, from which we found that 6 seems to be an appropriate choice of order. Figure 3.15 illustrates the squared cluster centers obtained from the k-pc algorithm when the order is chosen as 6. In Table 3.4, we convert the spectral estimation to the factor matrix B following the method with $\|\cdot\|_{(s)} = \|\cdot\|_{(r)} = \|\cdot\|_2$ and $\alpha = 2$. In addition, for each row of the matrix B, we find to which factor index (the same as the cluster index in Figure 3.15) the largest value (in bold) corresponds. We include these factor indices in the last column of Table 3.3, which can be viewed roughly as markings of groups of concomitant extremes. These 6 groups are in good accordance with the geographical context: most of the stations located along the same river are found in the same group, with the only exception of the 4 stations along the Mississippi River. The further division of these 4 stations into 2 groups may be easily justified by the large geographical

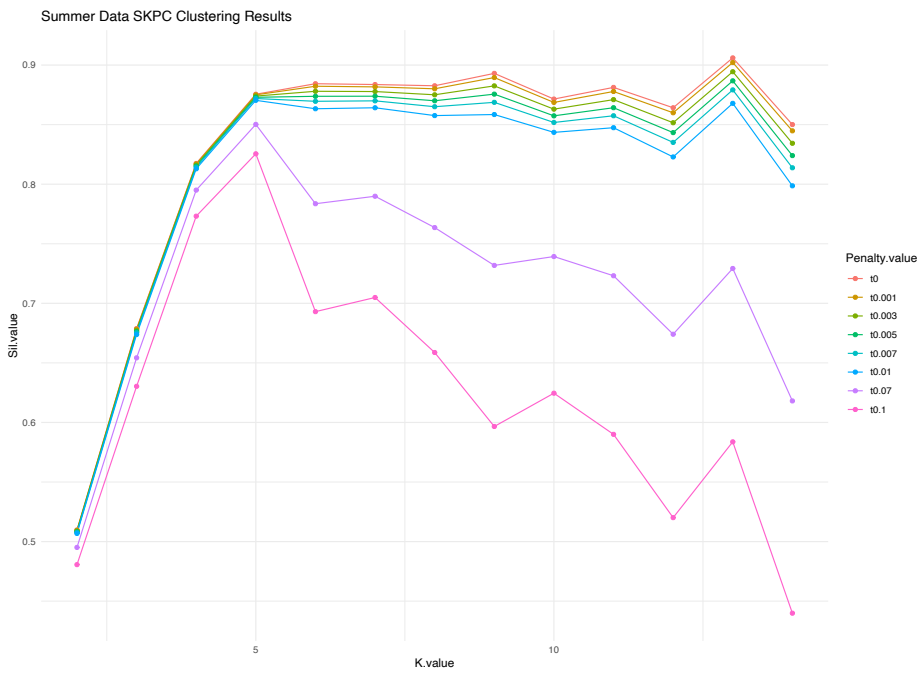


Figure 3.7: Air Pollution Example Summer-SK-PC Clustering

Table 3.3: River Discharge Stations

Station Name	River Name	Factor (Cluster) Index
SALEM, OR	WILLAMETTE RIVER	4
PORTLAND, OR	WILLAMETTE RIVER	4
HARRISBURG, OR	WILLAMETTE RIVER	4
BELOW SPRAGUE RIVER NEAR CHILOQUIN, OR	WILLIAMSON RIVER	2
ST.PAUL, MN	MISSISSIPPI RIVER	1
AITKIN, MN	MISSISSIPPI RIVER	1
THEBES, IL	MISSISSIPPI RIVER	6
CHESTER, IL	MISSISSIPPI RIVER	6
GREEN ISLAND, NY	HUDSON RIVER	5
FORT EDWARD, NY	HUDSON RIVER	5
NORTH CREEK, NY	HUDSON RIVER	5
NEAR CARLISLE, SC	BROAD RIVER	3
NEAR BELL, GA	BROAD RIVER	3

distance between the 2 groups: one group located in MN and the other located in IL.

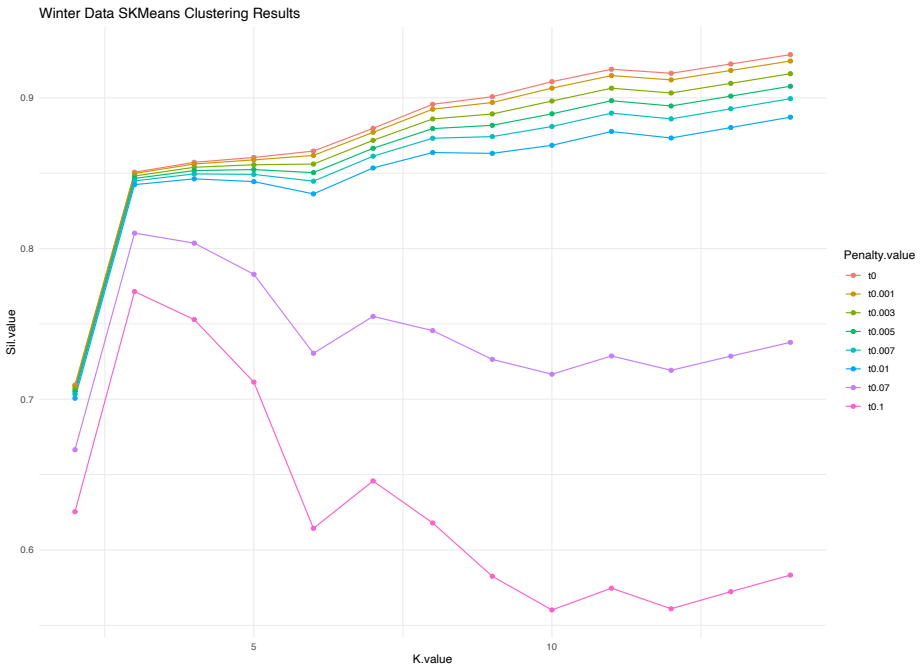


Figure 3.8: Air Pollution Example Winter-Spherical K Means Clustering

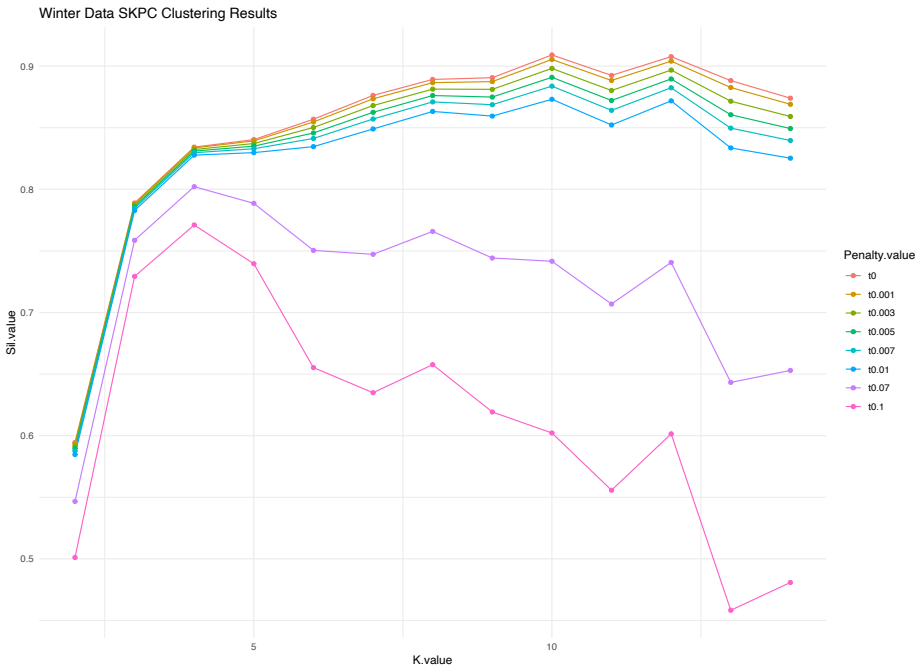


Figure 3.9: Air Pollution Example Winter-SK-PC Clustering

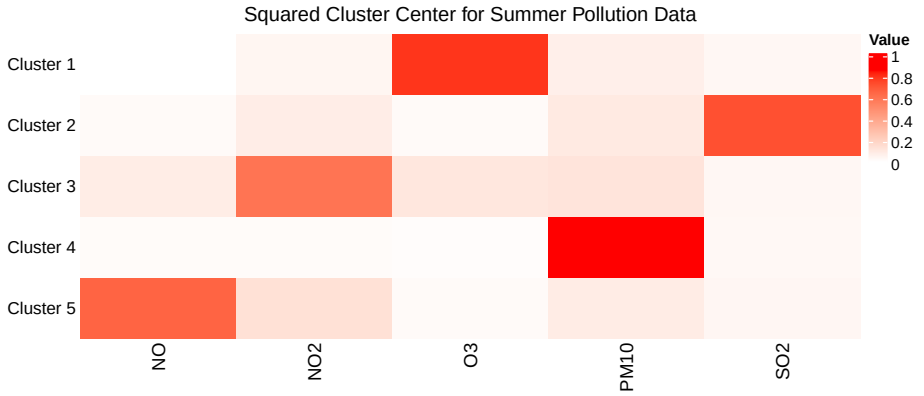


Figure 3.10: Air Pollution Example Winter-SK-PC Clustering

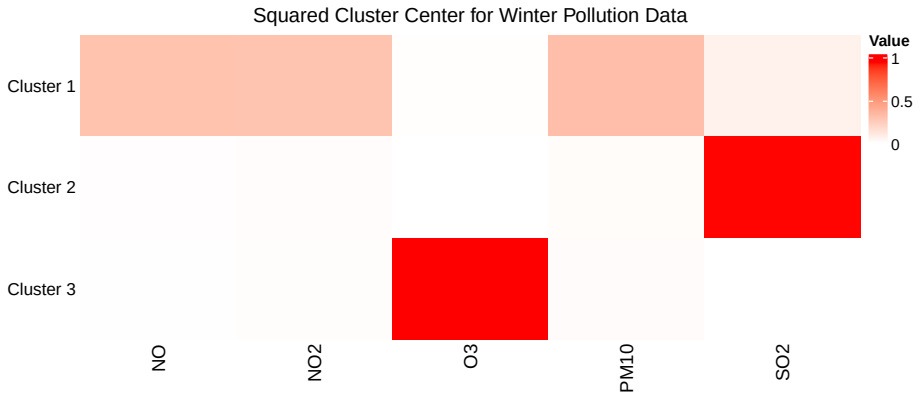


Figure 3.11: Air Pollution Example Winter-SK-PC Clustering

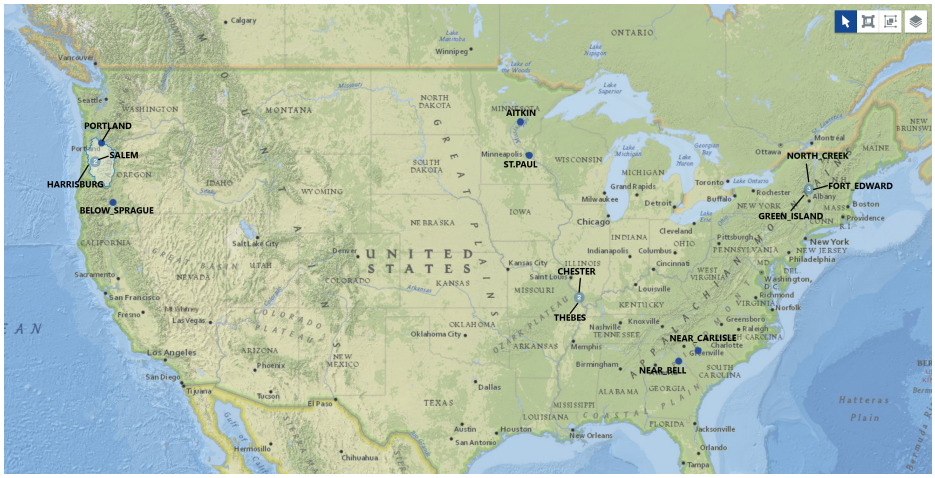


Figure 3.12: 13 River Discharge Stations

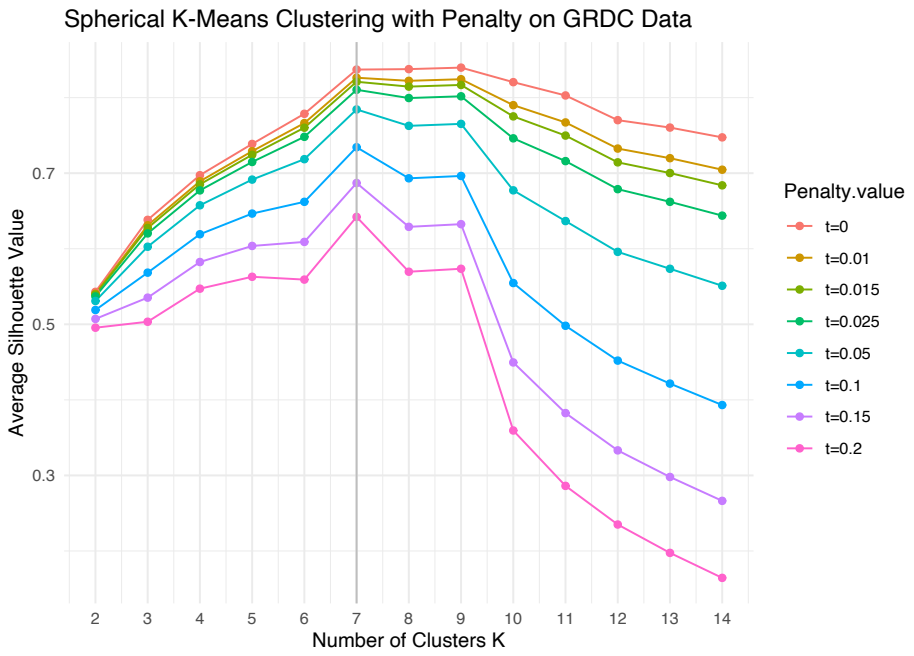


Figure 3.13: Water Discharge Example Spherical K Means Clustering

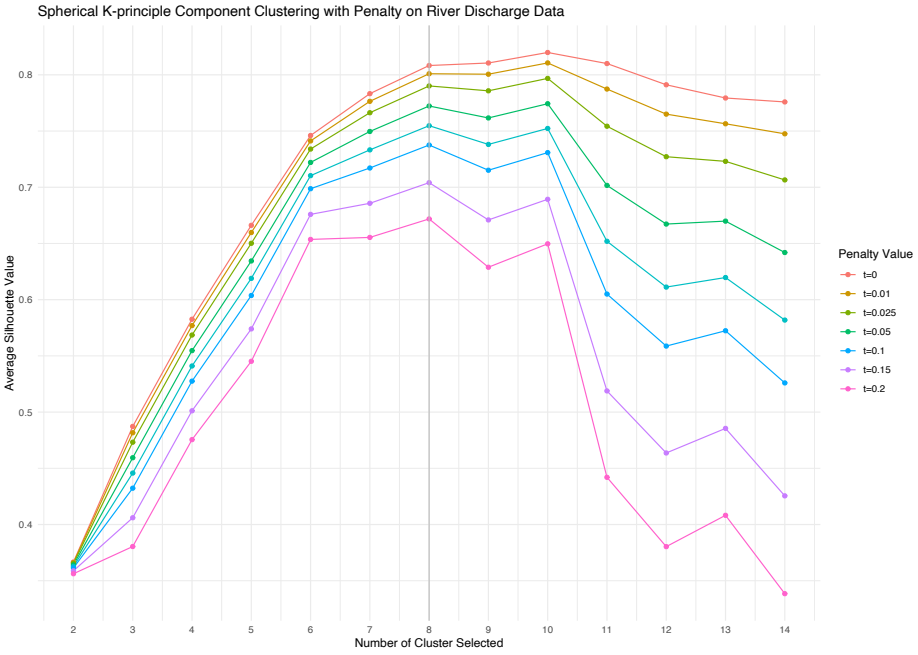


Figure 3.14: River Discharge Example SK-PC Clustering

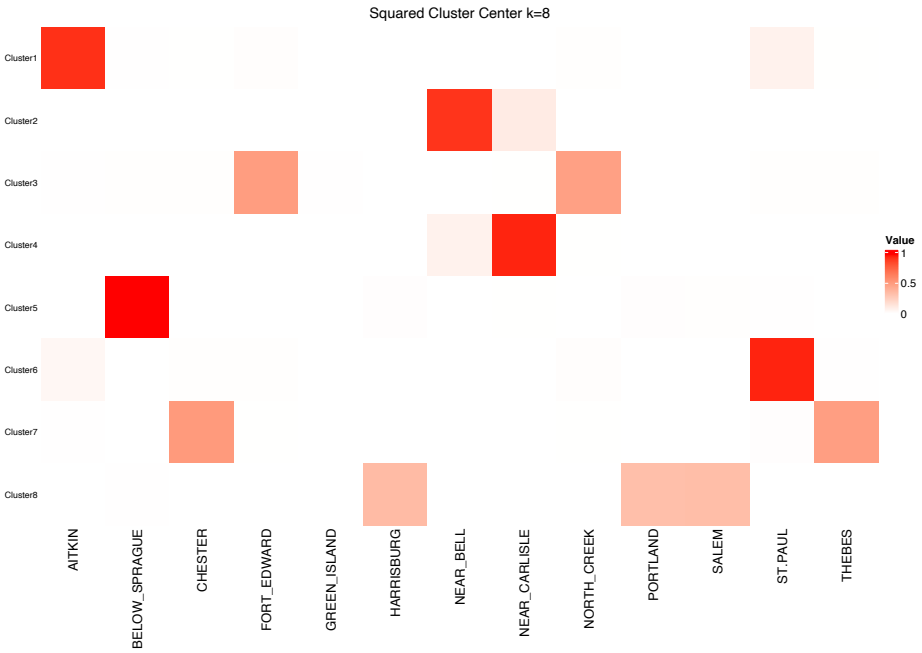


Figure 3.15: River Discharge Squared Cluster Center $k = 6$

Table 3.4: Estimated B for River Discharge Data

Factor	1	2	3	4	5	6
SALEM	0.15	0.26	0.11	**0.91**	0.19	0.18
PORTLAND	0.16	0.27	0.11	**0.91**	0.19	0.18
HARRISBURG	0.16	0.26	0.12	**0.91**	0.19	0.17
ST.PAUL	**0.88**	0.15	0.28	0.10	0.31	0.12
AITKIN	**0.91**	0.12	0.23	0.13	0.29	0.12
THEBES	0.28	0.15	**0.88**	0.11	0.31	0.16
CHESTER	0.29	0.15	**0.88**	0.10	0.30	0.16
BELOW_SPRAGUE	0.22	**0.87**	0.15	0.25	0.28	0.15
GREEN_ISLAND	0.41	0.26	0.28	0.32	**0.69**	0.35
FORT_EDWARD	0.31	0.12	0.18	0.16	**0.89**	0.17
NORTH_CREEK	0.30	0.12	0.17	0.16	**0.90**	0.18
NEAR_CARLISLE	0.15	0.17	0.15	0.19	0.22	**0.92**
NEAR_BELL	0.16	0.16	0.14	0.19	0.23	**0.92**

3.4.3 Discussion of Choice of Data Scale

Selecting an appropriate threshold based on extreme values in real-world data is a challenging problem, particularly as it can influence the optimal order selection when applying spherical clustering techniques to estimate linear factor models. To illustrate this, we use a summer air pollution dataset and explore different thresholds set at the 10%, 15%, and 20% largest l_2-norm values. For each threshold, we plot the range of Penalized ASW (Average Silhouette Width) scores for k values ranging from 1 to 12 and we use the SK-PC method. The optimal k is determined by analyzing the bending behavior of the corresponding Penalized ASW curves, and the resulting $B^\top$ matrix is computed based on this optimal k. This approach demonstrates how the choice of threshold affects the clustering and estimation outcomes.

The results in Figure 3.16, 3.18, and 3.20 showing that the proposed penalization has a significant bias correction effect on the selection of the optimal k. Furthermore, the results across different proportions as shown in Figure 3.17, 3.19, and 3.21 reveals that using a larger subset of data tends to yield smaller optimal k values, whereas using a smaller proportion leads to larger optimal k values.

In summary, the proportion of data selected based on the greatest norm can significantly influence the determination of the optimal order number k when employing spherical clustering techniques. A smaller proportion tends to result in larger optimal k values, whereas including too many observations in the analysis may obscure convergence and hinder clear clustering outcomes. Overall, selecting 10% of the data appears to be a reasonable choice in most cases.

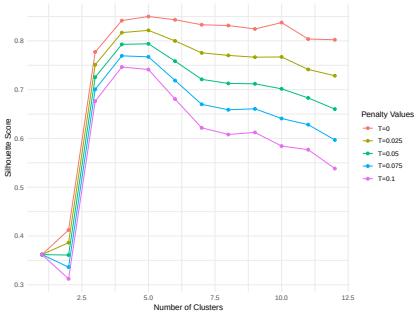


Figure 3.16: Penalized ASW Curves for Summer Air Pollution Data (10%)

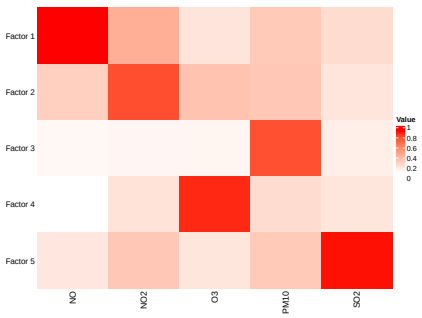


Figure 3.17: Estimated $B^\top$ (10%)

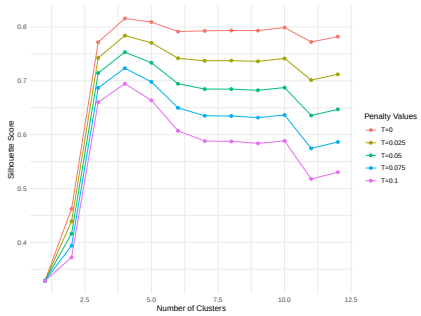


Figure 3.18: Penalized ASW Curves for Summer Air Pollution Data (15%)

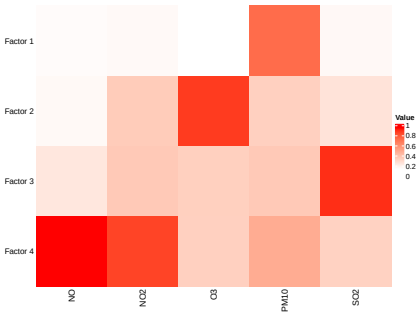


Figure 3.19: Estimated $B^\top$ (15%)

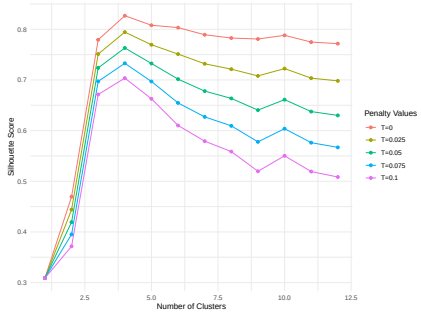


Figure 3.20: Penalized ASW Curves for Summer Air Pollution Data (20%)

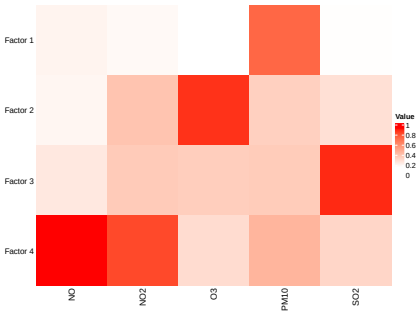


Figure 3.21: Estimated $B^\top$ (20%)

Chapter 4

Detection of Groups of Concomitant Extremes via Clustering

4.1 Introduction of Detection of Groups of Concomitant Extremes

The detection and analysis of concomitant extremes, where multiple extreme events occur together within the same period or in interconnected regions, are increasingly important in a world facing complex and interdependent risks. Concomitant extremes, such as simultaneous heatwaves and droughts, extreme rainfall with flooding, or concurrent market crashes, present unique challenges that traditional risk models often overlook. These events amplify risks and strain response mechanisms across critical sectors, including climate science, finance, public health, and infrastructure. Understanding and identifying these concurrent extreme events not only help in assessing their immediate impacts but also in developing strategies to mitigate cascading effects on ecosystems, economies, and communities.

A range of advanced methods have been developed to detect and analyze concomitant extremes. Threshold-based techniques, where variables are analyzed for simultaneous exceedance of defined critical levels, provide a straightforward way to identify occurrences of extremes. Copula models offer a sophisticated statistical approach for analyzing dependencies between variables, allowing researchers to quantify the probability of multiple extremes occurring together. Additionally, joint probability distributions for continuous data help evaluate the likelihood of multivariable extremes, while Multivariate Extreme

Value Theory (MEVT) extends traditional extreme value analysis to multiple variables, focusing on the tails of joint distributions where these extremes are more likely to co-occur. Each method provides unique insights into the nature and frequency of concomitant extremes, laying the groundwork for understanding their broader impacts.

Beyond detection, analyzing the development and evolution of concomitant extremes provides crucial insights into their underlying drivers and patterns. By employing trend analysis, time series methods, and machine learning models, researchers can uncover temporal patterns, cycles, and even emerging risks associated with these extreme events. Machine learning methods, in particular, excel at handling high-dimensional data and can reveal complex dependencies that may not be immediately apparent in traditional analysis. Event attribution techniques further enable scientists to link these extremes to specific causal mechanisms, such as atmospheric conditions, socioeconomic stressors, or even climate change, providing a comprehensive understanding of both the origins and potential future trajectories of these events.

The applications of detecting and analyzing concomitant extremes are far-reaching and essential for effective risk management and policy planning. In the insurance and finance sectors, quantifying the likelihood and potential impact of simultaneous extreme events aids in creating robust risk assessment models that more accurately reflect real-world conditions. For policymakers and urban planners, insights into the development of concomitant extremes support the creation of targeted adaptation strategies to increase resilience against climate change impacts and other risks. Moreover, understanding the compounded effects of concurrent extremes allows for more effective preparedness and response strategies across public health, infrastructure, and environmental management. By advancing our knowledge of concomitant extremes, we can make informed decisions to safeguard vital systems and improve resilience against multifaceted and interdependent risks in an increasingly complex world.

4.2 Spherical Sparse K Principle Component Clustering

4.2.1 Settings and definitions for Concomitant Extremes

In this chapter, follow the multivariate extreme setting. After standardizing the marginals, assume the random vector $\mathbf{Y}$ defined in (1.5) has spectral measure S

as defined in (1.6). Suppose $(Z_1, \cdots Z_d)$ follows the distribution S embedded in $\mathbb{R}^d$. The marginal standardization ensures that:

$$E(Z_1) = \cdots = E(Z_d) =: \mu > 0. \tag{4.1}$$

To describe the dependence structure of $\mathbf{Y}$, the pairwise dependence coefficient is commonly used and it is defined as follows: the tail dependence coefficient of Y_i and Y_j is

$$\chi_{ij} = \lim_{t\to\infty} pr(Y_j > t | Y_i > t) = \frac{1}{\mu} E(Z_i \wedge Z_j) \in [0, 1].$$

Here $\mu \in [1/d, 1/\sqrt{d}]$, and the necessary and sufficient conditions of the χ_{ij} attains its boundary are clearly stated in Lemma 1 from Fomichov and Ivanovs, 2023.

In high dimensional scenarios where d is large, let $I \in \{1, \cdots, d\}$ be a nonempty set of indices such that $P(X_i > 0,\ for\ all\ i \in I, X_j = 0\ for\ all\ j \notin I) > 0$, the corresponding faces are defined as:

Definition 4

$$F_I = \{x \in \mathbb{R}^d_+ : x_j = 0\ for\ j \notin I\} \tag{4.2}$$

In this proposal, there is a universal assumption for the faces: There exists $2 \leq k \leq d$ and a partition $(I_1, \cdots, I_k)$ of the index set $\{1, \cdots, d\}$ satisfying $pr(x \in F_{I_1} \cup \cdots \cup F_{I_k}) = 1$. Without loss of generality, assume that for all $1 \leq i < j \leq k$, the indices in I_i are smaller than the indices in I_j.

Figure 4.1 shows a 3-dimensional case satisfying assumption 4.2.1. The data is generated from a mixture of two spherical Dirichlet distributions with a mean approximately equal to 0.4 in each of the three directions. It is clear from the figure that there exist two faces $I_1 = \{1\}$ and $I_2 = \{2, 3\}$ with corresponding dimensions $d_1 = 1$ and $d_2 = 2$. The two centers are $c_1 = (0.7, 0.7, 0)$ and $c_2 = (0, 0, 1)$, and the proportions of the data around the centers are about $p1 = 57.6\%$ and $p_2 = 42.4\%$. Figure 4.1 visualizes a three-dimensional example, and the detection of concomitant extremes is to propose faces $\hat{I}_1$ and $\hat{I}_2$ based on data which hopefully coincide with $I_1 = \{1\}$ and $I_2 = \{2, 3\}$.

Under the assumption 4.2.1 and (4.1), the cross-moment matrices could be defined as:

Definition 5 *Let $X_I \in \mathbb{S}_+^{|I|-1}$ with $\mu_I = \mu/p_I$, the cross-moment matrices are given by*

$$\Sigma_I = E(X_I X_I^T),\ \Sigma = \Sigma_I = E(XX^T) = diag(p_{I_1}\Sigma_{I_1}, \cdots, p_{I_k}\Sigma_{I_k}) \tag{4.3}$$

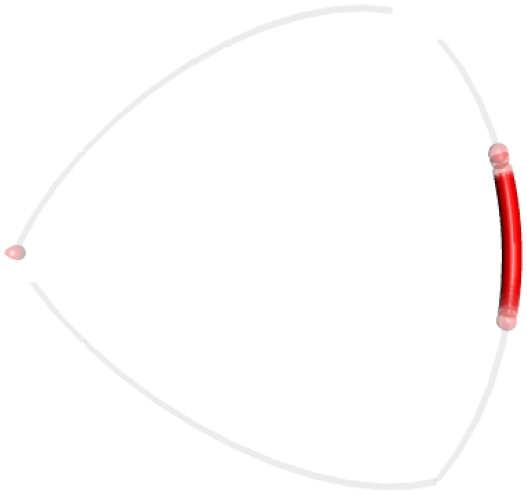


Figure 4.1: Mixture of Two Dirichlet Distribution

Every Σ_I is a nonnegative definite matrix with trace 1, since $\|X_I\|_2 = 1$. In the Supplementary Material of Fomichov and Ivanovs, 2023, it mentioned that Σ can be used to check sufficient conditions in the result.

4.2.2 Spherical Sparse K Principle Component Clustering Method

On the grounds of spherical k-pc clustering, we proposed an improved spherical sparse k-pc clustering by introducing the sparse principal component Zou et al., 2006. The algorithm 4.2.2 gives a single iteration similar to algorithm 4.2.2.

Algorithm

- **Input:** the sample $\theta_1 \cdots \theta_n \in \mathbb{S}_+^{d-1}$ and current centroids $x_1, \cdots, x_k \in \mathbb{S}_+^{d-1}$

- **Same as Step 1- Step 3 in Algorithm 4.2.2**
- **Step 4:** For $i = 1$ to $i = k$, calculate $\Sigma_i = (1/n)\sum_{\mu=1}^{n}(\theta_\mu \theta_\mu^T 1_{\{g_\mu = i\}})$, and find the sparse principal eigenvector $\hat{x}_i \in \mathbb{S}_+^{d-1}$ of Σ_i with penalty parameter λ_n defined in 6. (See Zou et al., 2006
- **Output:** new centroids $\hat{x}_1, \cdots \hat{x}_k \in \mathbb{S}_+^{d-1}$ and the old value v.

Definition 6 *Suppose* $\mathbf{X}$ *is a random vector in* $\mathbb{S}_+^{d-1}$ *and* $\Sigma = E(X^T X)$. *The sparse principle eigen vector (See Zou et al., 2006)* $\hat{\mathbf{x}}$ *of* Σ *is defined as:*

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathbf{x}^T \Sigma \mathbf{x} + \lambda \|\mathbf{x}\| + \lambda_1 \|\mathbf{x}_1\|, \tag{4.4}$$

where λ_1 is the penalty parameter and λ is chosen to be a small positive number to overcome potential collinearity problems from Zou et al., 2006.

By introducing the first sparse principle component as a substitute of the center in the algorithm, it tends to shrink the prototype coordinate into a lower-dimension face (cf. (4.2.1)). In other words, in detecting concomitant faces, through encouraging sparsity, we have more decisive results in disassembling dimensions into lower-dimension components.

4.3 Data Example

4.3.1 Simulate Data

Data From Max-linear Model

In this section, a three-dimensional dataset is generated from the max-linear model in (2.1). We first generated 1000 data points with the greatest norm from 10000 simulations. The simulated data are from a max-linear model with centers $a_1 = (0, 1, 0)$, $a_2 = (0, 0, 1)$, and $a_3 = (1, 0, 0)$.

In the analysis, three clustering methods—spherical k-means, spherical k-pc, and spherical sparse k-pc—are utilized with the number of clusters set to $k = 3$. The prototypes of these clusters are presented in table 4.1. It is evident from the table that the spherical sparse k-pc method effectively pushes the cluster centers towards the boundaries. In contrast, the Spherical k-means and spherical k-pc methods do not achieve the same level of separation. This property of spherical sparse k-pc suggests that it can better detect faces in a high-dimensional scenario where some faces are in close proximity to each other.

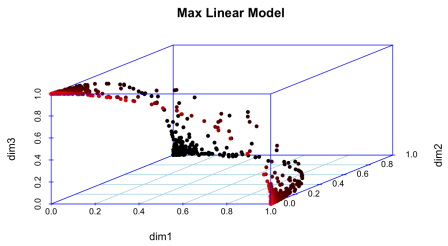


Figure 4.2: Simulated 3-dimensional data

Table 4.1: Optimal Centers

	S k-means	S k-pc	S Sparse k-pc
Center1	(0.055, 0.997, 0.058)	(0.053, 0.997, 0.055)	(0, 1, 0)
Center2	(0.057, 0.060, 0.997)	(0.055, 0.058, 0.997)	(0, 0, 1)
Center3	(0.996, 0.062, 0.068)	(0.996, 0.059, 0.064)	(1, 0, 0)

The enhanced ability of spherical sparse k-pc to distribute cluster centers near the boundaries indicates its potential to discern finer distinctions and capture more complex patterns in high-dimensional data, making it a promising choice for face detection and related tasks.

Data from a mixture of two Dirichlet Distribution

The following describes the process of generating a dataset from a two-mixture Dirichlet model. The generated data is normalized and has the same expectation in all dimensions. Given parameters:

- n: Number of samples to generate.
- I_1, I_2: The number of dimensions in the first and second Dirichlet distributions, respectively.
- α_1, α_2: Total concentration parameters for the two Dirichlet distributions.

The process involves the following steps:

1. Compute Per-Dimension α: The per-dimension concentration parameters for the two Dirichlet distributions are calculated as:

$$\alpha_{1i} = \frac{\alpha_1}{I_1}, \quad \alpha_{2i} = \frac{\alpha_2}{I_2}.$$

2. Calculate Mixture Probabilities (ϕ_1 and ϕ_2): Mixture probabilities are derived using the gamma function:

$$\phi_1 = \frac{\Gamma(\alpha_{1i} + 0.5)}{\Gamma(\alpha_{1i})} \cdot \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1 + 0.5)}, \quad \phi_2 = \frac{\Gamma(\alpha_{2i} + 0.5)}{\Gamma(\alpha_{2i})} \cdot \frac{\Gamma(\alpha_2)}{\Gamma(\alpha_2 + 0.5)}.$$

3. Solve for Mixture Probabilities (P_1, P_2): Using a linear system defined by:

$$\begin{bmatrix} \phi_1 & -\phi_2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

the probabilities P_1 and P_2 are determined. 4. Generate Samples: For each sample:

1. Perform a Bernoulli trial with probability P_1 to decide between the two Dirichlet distributions.

2. If the first Dirichlet is chosen:

   Sample from Dirichlet with parameter $\boldsymbol{\alpha} = [\alpha_{1i}, \ldots, \alpha_{1i}]$ (length I_1),

   and normalize to fit the dimensional requirements.

3. If the second Dirichlet is chosen:

   Sample from Dirichlet with parameter $\boldsymbol{\beta} = [\alpha_{2i}, \ldots, \alpha_{2i}]$ (length I_2),

   and normalize similarly.

4. Add small Gaussian noise to the unused dimensions to preserve consistency.

5. Normalize and Return Data: The resulting data matrix d is returned, where each row corresponds to a sample from the mixture distribution.

The generated data has the following properties:

- Normalization: Each sample is normalized, ensuring that the sum of values across all dimensions equals 1. This is a characteristic of the Dirichlet distribution.

- Equal Expectations: The expectation of each dimension is the same, reflecting the uniform concentration of the Dirichlet distributions in the mixture.

- Mixture Structure: The data exhibits properties of both components of the mixture, with overlap or distinctiveness depending on the parameters $I_1, I_2, \alpha_1, \alpha_2$.

This algorithm generates data that is particularly useful for modeling phenomena with normalized proportions and symmetric expectations across dimensions.

Here, we generated 1000 data with dimension equals 100 where $I_1 = 15$ and $I_2 = 20$, $\alpha_1 = 40$ and $\alpha_2 = 70$.

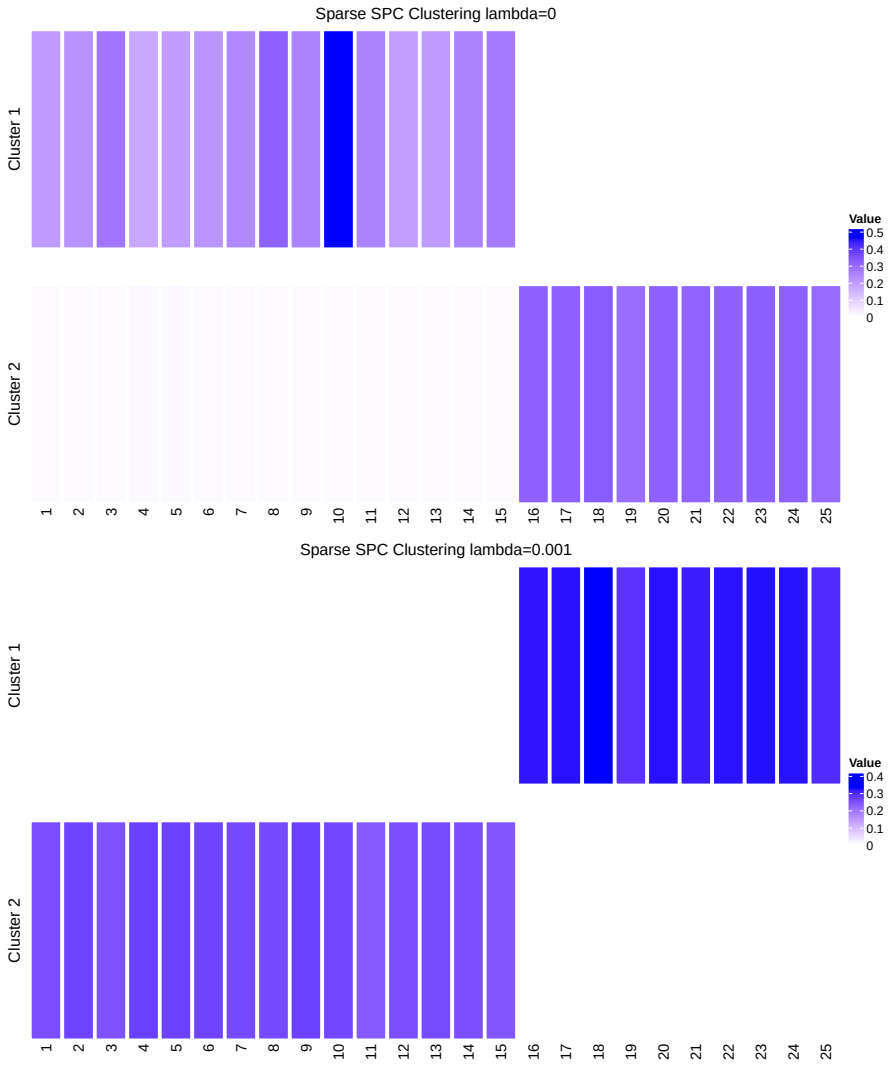


Figure 4.3: Simulated mixture of Dirichlet Sparse K-PC Clustering

Table 4.2 and Figure 4.3 together demonstrate the effectiveness of incorporating a LASSO penalty into the Spherical K-PC Clustering Method for analyzing a simulated mixture of Dirichlet data. The table provides quantitative

Table 4.2: Comparison of Prototypes from Spherical K-PC Clustering and Sparse Spherical K-PC Clustering

	SKPC-center1	**SKPC-center2**	**S-SKPC-center1**	**S-SKPC-center2**
1	0.20045209	0.01196792	0.0000000	0.2535923
2	0.21998864	0.01236777	0.0000000	0.2665177
3	0.27814564	0.01108584	0.0000000	0.2473707
4	0.17449421	0.01296781	0.0000000	0.2724081
5	0.19767500	0.01268790	0.0000000	0.2699303
6	0.21572340	0.01237430	0.0000000	0.2661480
7	0.23838061	0.01190837	0.0000000	0.2595561
8	0.32160602	0.01126395	0.0000000	0.2597794
9	0.25214388	0.01224915	0.0000000	0.2698147
10	0.46679981	0.01019717	0.0000000	0.2621016
11	0.25248716	0.01088808	0.0000000	0.2384005
12	0.19270288	0.01198484	0.0000000	0.2525055
13	0.20036509	0.01213581	0.0000000	0.2577425
14	0.25101390	0.01146864	0.0000000	0.2519010
15	0.26874840	0.01093459	0.0000000	0.2423480
16	0.00016871	0.31897016	0.3195822	0.0000000
17	0.00021999	0.32098217	0.3218231	0.0000000
18	0.00016383	0.32946725	0.3311275	0.0000000
19	0.00013747	0.29215165	0.2896996	0.0000000
20	0.00026310	0.32087971	0.3217140	0.0000000
21	0.00017207	0.31334197	0.3131613	0.0000000
22	0.00021140	0.31917314	0.3203272	0.0000000
23	0.00017502	0.32270064	0.3234584	0.0000000
24	0.00020595	0.32053280	0.3214657	0.0000000
25	0.00018639	0.29890995	0.2976255	0.0000000

insights into how feature weights are distributed across clusters, while the figure offers a visual representation of the feature importance before and after applying regularization. Without the LASSO penalty, the feature contributions in both clusters appear more evenly distributed, as seen in the table's entries for $\lambda = 0$, where many feature values are non-zero. Similarly, in the top panel of the figure, the bars for both clusters are dense and uniformly intense across most dimensions. This lack of sparsity makes it difficult to identify which features are most relevant for distinguishing the clusters, potentially leading to noise and overfitting in the clustering results.

By introducing the LASSO penalty ($\lambda = 0.001$), the method enhances sparsity by shrinking less significant features toward zero. This effect is evident in the table, where many entries for $\lambda = 0.001$ are close to zero, particularly in less relevant dimensions. The corresponding bottom panel of Figure 4.3 visually confirms this, showing lighter bars in irrelevant dimensions and darker bars in dimensions with significant contributions. This regularization effectively reduces the complexity of the parameter space, enabling the method to focus on the most critical features for clustering. The sparsity achieved through the LASSO penalty ensures that irrelevant features do not influence the clustering results, improving the robustness and interpretability of the model.

The combined results from the table and figure demonstrate the ability of the LASSO-regularized Spherical K-PC Clustering Method to identify the distinct characteristics of the two clusters more clearly. The sparsity induced by the penalty allows the method to emphasize the most important dimensions, isolating the unique features of each cluster. This leads to a clearer separation between the two clusters, as the method can better capture their underlying structure. Overall, the addition of the LASSO penalty enhances both the clustering performance and the interpretability of the results, providing a more effective tool for analyzing high-dimensional data with sparse and structured patterns.

4.3.2 Real Data Example

4.3.3 Air Pollution Data

From Figure 4.4, the results for summer air pollution data show the clustering structure before and after applying the sparse penalty ($\lambda = 0.015$). Without the sparse penalty ($\lambda = 0$), the top heatmap illustrates that all five clusters rely on multiple pollutants (O_3, NO_2, NO, SO_2, PM_{10}) to varying degrees, as indicated by the spread of red and orange shades across all dimensions. This lack of sparsity makes it harder to pinpoint which pollutants dominate specific clusters, leading to less interpretable results.

When the sparse penalty is applied ($\lambda = 0.015$), the bottom heatmap demonstrates a clear reduction in the number of dominant pollutants for each cluster. For example, PM_{10} becomes the primary driver of Cluster 5, while NO_2 and NO dominate other clusters. This sparsity simplifies the interpretation, as each cluster's defining characteristics become clearer. The summer results indicate that pollutant concentrations are distributed across more dimensions under the non-sparse case, but the sparsity-enforced structure emphasizes critical pollutants associated with each cluster.

Similarly, the winter air pollution data shows clustering results in Figure 4.5 with and without sparsity ($\lambda = 0.025$). Without regularization ($\lambda = 0$), the top heatmap reveals that pollutants such as NO, SO_2, and PM_{10} have substantial contributions across clusters. However, the widespread presence of non-zero values across dimensions again makes it challenging to isolate the pollutants most relevant to specific clusters.

With the sparse penalty ($\lambda = 0.025$), the bottom heatmap reveals a more structured and interpretable clustering pattern. Clusters are now associated with fewer dominant pollutants, such as O_3 driving Cluster 1 and PM_{10} dominating another. This result indicates that sparsity aids in highlighting seasonally relevant pollutants, particularly during winter, where different weather and atmospheric conditions may alter the significance of specific pollutants.

4.3.4 Air Pollution Data

The application of the Sparse Spherical K-PC Clustering method demonstrates clear seasonal differences in the clustering structure. In summer, pollutants like NO_2 and PM_{10} appear to have broader significance across clusters, while in winter, O_3 and PM_{10} emerge as more dominant pollutants for specific clusters under sparse conditions. The difference in clustering patterns reflects the seasonal variability in pollution sources and atmospheric conditions, such as increased photochemical reactions in summer and the prominence of heating emissions in winter.

Additionally, the sparsity-enforced structure enhances interpretability in both cases, reducing dimensional noise and focusing on the key pollutants associated with each season. These results highlight the utility of sparse clustering for uncovering meaningful seasonal patterns in complex, high-dimensional environmental datasets.

4.3.5 Daily River Discharge Data

Figure 4.6 demonstrates the effect of introducing a sparse penalty ($\lambda = 0.007$) in the Spherical K-PC Clustering Method compared to the case without regularization ($\lambda = 0.0$). The sparse penalty pushes many feature contributions to zero, as seen in the top heatmap, where most cells are light-colored. This sparsity structure allows the method to focus on the most relevant features for clustering, highlighting the sites with significant contributions in each cluster while filtering out less important ones. In contrast, the clustering without the sparse penalty ($\lambda = 0.0$) results in broader, more uniformly distributed feature contributions, making it harder to interpret the key drivers of each cluster.

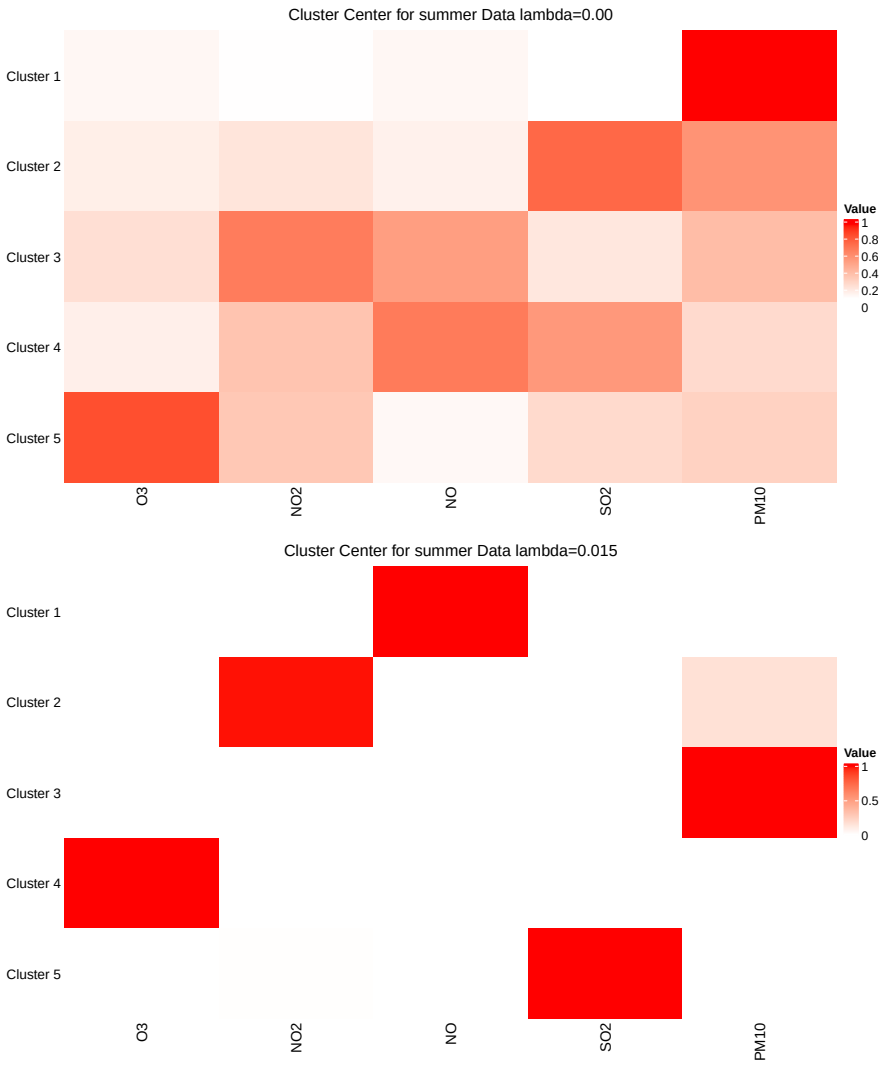


Figure 4.4: Sparse Spherical K-Principle Component Clustering on Summer Air Pollution Data

The sparsity induced by the LASSO penalty not only simplifies the clustering structure but also helps identify distinct "faces" or dominant groups within the dataset. These faces represent combinations of river sites that have similar patterns and behaviors, forming meaningful groupings. For instance, certain clusters in the sparse penalty case are dominated by a few key sites like "CHESTER" or "BELOW_SPRAGUE," making it easier to distinguish between groups based on their unique characteristics.

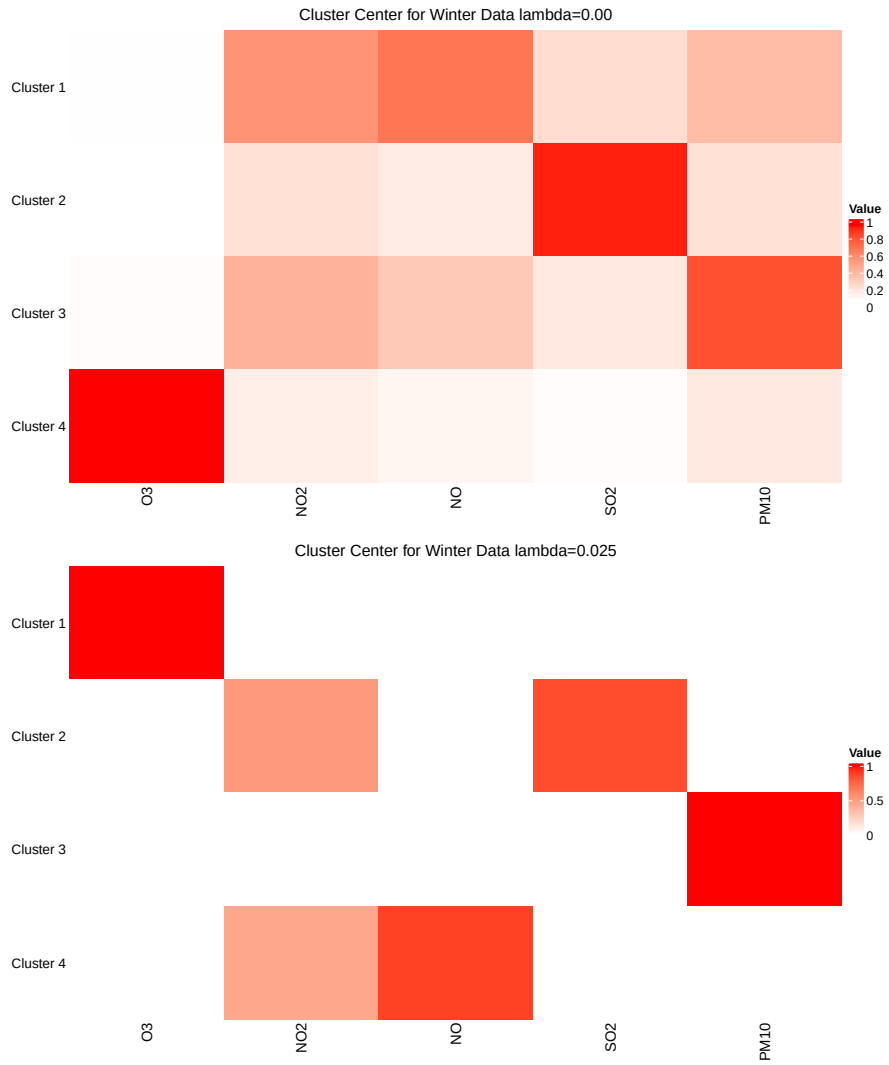


Figure 4.5: Sparse Spherical K-Principle Component Clustering on Winter Air Pollution Data

Importantly, these faces are meaningful in real-world scenarios as they correspond to geographical locations with similar hydrological behaviors. The identified clusters reflect the physical proximity or shared characteristics of the river sites, such as their response to extreme weather events or seasonal patterns. This alignment between the sparse clustering results and real-world locations underscores the practical utility of incorporating sparsity, enabling the model to provide insights that are both interpretable and actionable.

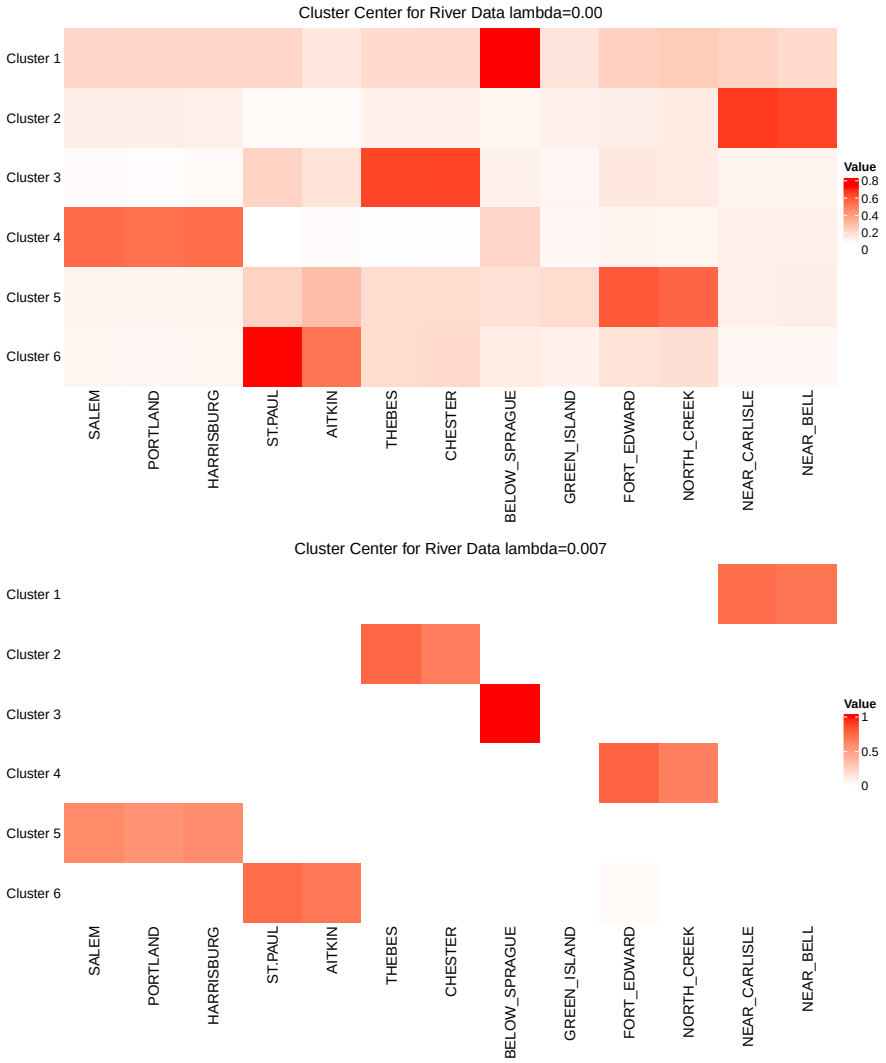


Figure 4.6: Sparse Spherical K-Principle Component Clustering on Daily River Discharge Data

In summary, the introduction of a sparse penalty in the Sparse K-PC Clustering Method significantly enhances the sparsity of the clustering structure. By applying the LASSO regularization, the method effectively pushes less significant site contributions toward zero, reducing noise and emphasizing the most important features. This sparsity not only simplifies the parameter space but also allows the method to isolate the dominant patterns in the data, making the clusters more distinct and interpretable.

The sparse penalty plays a crucial role in identifying the key "faces" of the clustering solution. These faces represent the most influential river sites that define each cluster. The sparsity induced by the penalty ensures that only a subset of the features contributes meaningfully to the clustering process, highlighting the primary sites that differentiate one cluster from another. Without this regularization, the clustering structure would remain more complex, with non-essential features masking the true underlying patterns.

Importantly, these identified faces are meaningful in real-world contexts, as they correspond to the geographic locations of the river sites. The clusters reveal patterns in river discharge data that align with the physical and environmental characteristics of the locations. This correspondence ensures that the clustering results are not only mathematically robust but also relevant for practical applications, such as hydrological modeling and water resource management. The sparse penalty thus provides a powerful tool for uncovering interpretable and actionable insights in complex datasets.

Chapter 5

Conclusion

In this thesis, we explored the application of linear factor models and spherical clustering techniques to analyze multivariate extremes, with a particular focus on enhancing estimation and order selection methods. Beginning with an overview of extreme value analysis, we provided a foundational understanding of univariate and multivariate extreme value theory, which guided our approach to clustering methods in extreme datasets.

We introduced and evaluated the use of spherical k-means and spherical k-principal component clustering in estimating linear factor models, showing their effectiveness in identifying patterns among extreme values. Additionally, we proposed a penalized silhouette method for order selection, addressing the need for optimal cluster determination in high-dimensional data. This method proved valuable in enhancing the robustness of clustering outcomes by balancing precision and computational efficiency.

Furthermore, we extended our analysis to sparse spherical clustering methods, demonstrating their utility in detecting groups of concomitant extremes. This approach was particularly effective for identifying relevant clusters in scenarios where extreme values are highly interdependent, providing a practical solution for clustering in complex, high-dimensional datasets.

Overall, this research contributes to the field of multivariate extreme value analysis by developing and refining clustering techniques suitable for extreme data scenarios. Our findings underscore the importance of thoughtful model estimation, order selection, and the integration of sparse clustering approaches, providing a framework that can be applied in various fields such as risk management, environmental science, and finance. Future work may explore additional refinements to the proposed methods and consider their applications in real-time, large-scale data environments.

Appendix A

The R codes that implement the simulation and real data studies can be found at https://github.com/SyuanD/SphCluster.git.

Bibliography

Bader, B., Yan, J., & Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate.

Balkema, A. A., & De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, *2*(5), 792–804.

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. L. (2006). *Statistics of extremes: Theory and applications*. John Wiley & Sons.

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

Bottolo, L., Consonni, G., Dellaportas, P., & Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes*, *6*, 25–47.

Chen, X. (2014). Extreme value distribution and peak factor of crosswind response of flexible structures with nonlinear aeroelastic effect. *Journal of Structural Engineering*, *140*(12), 04014091.

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

Coles, S. G., & Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(2), 377–392. Retrieved April 21, 2023, from http://www.jstor.org/stable/2345748

Coles, S. G., & Tawn, J. A. (1994). Statistical methods for multivariate extremes: An application to structural design. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*(1), 1–31.

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

Deng, S., Tang, H., & Bai, S. (2024). On estimation and order selection for multivariate extremes via clustering. https://arxiv.org/abs/2406.14535

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

Einmahl, J. H., Kiriliouk, A., Krajina, A., & Segers, J. (2016). An m-estimator of spatial tail dependence. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 275–298.

Einmahl, J. H., Kiriliouk, A., & Segers, J. (2018). A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes*, *21*, 205–233.

Einmahl, J. H., Krajina, A., & Segers, J. (2012). An m-estimator for tail dependence in arbitrary dimensions.

Einmahl, J. H., Piterbarg, V. I., & De Haan, L. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, *29*(5), 1401–1423.

Engeland, K., Hisdal, H., & Frigessi, A. (2004). Practical extreme value modelling of hydrological floods and droughts: A case study. *Extremes*, *7*, 5–30.

Engelke, S., & Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, *8*, 241–270.

Falk, M., & Marohn, F. (1993). Von mises conditions revisited. *The Annals of Probability*, 1310–1328.

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. (2023). Spherical clustering in detection of groups of concomitant extremes. *Biometrika*, *110*(1), 135–153.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Math. Polon.*, *6*, 93–116.

Galvin, F., & Shore, S. (1984). Completeness in semimetric spaces. *Pacific Journal of Mathematics*, *113*(1), 67–75.

German Federal Institute of Hydrology. (n.d.). Global runoff data centre (grdc) portal.

Gissibl, N., & Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs.

Gnecco, N., Meinshausen, N., Peters, J., & Engelke, S. (2021). Causal discovery in heavy-tailed models. *The Annals of Statistics*, *49*(3), 1755–1778.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, *44*(3), 423–453. Retrieved April 20, 2023, from http://www.jstor.org/stable/1968974

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

Heffernan, J. E., & Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *66*(3), 497–546.

Hornik, K., Feinerer, I., Kober, M., & Buchta, C. (2012a). Spherical *k*-means clustering. *Journal of Statistical Software*, *50*(10), 1–22. https://doi.org/10.18637/jss.v050.i10

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

Hruschka, E. R., de Castro, L. N., & Campello, R. J. (2004). Evolutionary algorithms for clustering gene-expression data. *Fourth IEEE International Conference on Data Mining (ICDM'04)*, 403–406.

Hussain, S. I., & Li, S. (2015). Modeling the distribution of extreme returns in the chinese stock market. *Journal of international Financial Markets, institutions and money*, *34*, 263–276.

Janßen, A., & Wan, P. (2020a). *k*-means clustering of extremes. *Electronic Journal of Statistics*, *14*(1), 1211–1233. https://doi.org/10.1214/20-EJS1689

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

Klüppelberg, C., & Lauritzen, S. (2019). Bayesian networks for max-linear models. *Network Science: An Aerial View*, 79–97.

Medina, M. A., Davis, R. A., & Samorodnitsky, G. (2021). Spectral learning of multivariate extremes. *arXiv preprint arXiv:2111.07799*.

Moriya, T., Roth, H. R., Nakamura, S., Oda, H., Nagara, K., Oda, M., & Mori, K. (2018). Unsupervised pathology image segmentation using representation learning with spherical k-means. *Medical Imaging 2018: Digital Pathology*, *10581*, 278–284.

Moussa, M., & Măndoiu, I. I. (2018). Single cell rna-seq data clustering using tf-idf based methods. *BMC genomics*, *19*, 31–45.

Rényi, A. (1963). On stable sequences of events. *Sankhyā: The Indian Journal of Statistics, Series A*, 293–302.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tunali, V., Bilgin, T., & Camurcu, A. (2016). An improved clustering algorithm for text mining: Multi-cluster spherical k-means. *International Arab Journal of Information Technology (IAJIT)*, *13*(1).

Von Mises, R. (1936). La distribution de la plus grande de n valuers. *Rev. math. Union interbalcanique*, *1*, 141–160.

Wan, P., & Davis, R. A. (2019). Threshold selection for multivariate heavy-tailed data. *Extremes*, *22*(1), 131–166.

Wilson, W. A. (1931). On semi-metric spaces. *American Journal of Mathematics*, *53*(2), 361–373.

Yuen, R., & Stoev, S. (2014). Crps m-estimation for max-stable models. *Extremes*, *17*, 387–410.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, *15*(2), 265–286.