Statistical learning methods for dimension reduction in multivariate extremes

Matthew Pawley

Thesis Formulation Report

September 2021

Supervised by Dr Christian Rohrbeck and Dr Evangelos Evangelou

Department of Mathematical Sciences

University of Bath

Abstract

Extreme value theory provides a rigorous mathematical framework for modelling the tail behaviour of multivariate random variables, which is of interest in many application areas. Typically, the marginal distributions and extremal dependence structure are modelled separately. The extremal dependence structure describes the tail dependence between components and is characterised by a measure, called the angular measure, which must be estimated based on a small number of observed extreme events. Traditional methods for modelling the angular measure are limited to small or moderate dimensions. Recently, methods from unsupervised learning, such as clustering and principal components analysis (PCA), have been adapted to extremes and are better-equipped to handle high-dimensional settings. This report comprises a critical review of such methods, with a focus on the framework for extremal PCA developed by Cooley and Thibaud (2019) and the algorithm for generating hazard event sets proposed by Rohrbeck and Cooley (2021). After applying these methods to analyse extreme rainfall in France, we outline avenues for future research.

Responsible Research & Innovation

This project concerns the development of statistical learning methodology for estimating the dependence structure in high-dimensional extremes. In particular, we aim to improve methods for simulating extreme events and consider an example of modelling extreme rainfall. Synthetic event sets may be used by practitioners in order to assess and mitigate risk. For example, the height of a new dam may be chosen so that it protects against all water levels that it is likely to experience within its projected life span of, say, 100 years. The importance of advancing modelling for climate extremes is demonstrated by the devestating human and environmental impacts of recent record-breaking flood and wildfire events.

Our methods can be applied more broadly beyond climate extremes in a wide variety of areas, such as engineering, oceaography and finance. However, our modelling assumptions cannot hope to be valid, or even be entirely realistic, in all settings that may be considered. For example, asymptotic independence is a degenerate case in the framework that underlies our methods. In many settings, this will drastically bias the estimation of probabilities in the joint tail, leading to an under/overestimation of the risk of certain events. Thus, the methods should be used primarily as an exploratory tool. As mathematicians, we must guard against incentives to present our methods as practically applicable in settings or ways that they are not. This is especially true given the potentially severe consequences of the events considered in extremes. Most consumers of our research will be other academics or sophisticated practitioners, but the outcomes may occassionally be presented to broader audiences. In any case, we should endeavour to always make clear any assumptions on which our methods are based.

In this report, we illustrate and test our methods using an example dataset from the field of meteorology. There are no ethical considerations concerning the use of this dataset. The data was initially collected by Météo-France, the French meteorological service, and is freely available from the website of Dr Philippe Naveau, an academic at Laboratoire des Sciences du Climat et l'Environnement (LSCE) in France.

Contents

Appendices

List of Figures

1
Introduction

In many applications, such as environmental science and finance, we are interested in analysing the tail behaviour of a multivariate random variable in order to assess the risk of certain extreme events[1]. This calls for a multivariate analysis using tools from multivariate extreme value theory, central to which is the notion of extremal dependence, i.e. the tail dependence between components of a random vector. For example, practitioners need to be aware if several locations are likely to be jointly impacted by heavy rainfall, so that they can take action to mitigate the risk of large-scale flooding. The need to quantify extremal dependence is core to statistical techniques for multivariate extremes, but it is inherently challenging in high-dimensional settings.

Mathematically speaking, the extremal dependence structure is characterised by a high-dimensional measure, called the angular measure. Estimating distributions in high dimensions is challenging; in the extremes setting it becomes harder still because only a small fraction of the data contains an informative signal for the distributional tail, limiting our effective sample size. Classical techniques, such as semi- or fully-parametric models, are generally too restrictive, suffer the curse of dimensionality, or are computationally infeasible (Gumbel 1960; Tawn 1988; Hüsler and Reiss 1989; Wackernagel 1995; Boldi and Davison 2007; de Carvalho and Davison 2014; Hanson et al. 2017). Instead, researchers are exploring how tools from statistical learning, such as clustering and principal component analysis (PCA), can be adapted for extremes (Bernard et al. 2013; Chautru 2015; Cooley and Thibaud 2019; Fomichov and Ivanovs 2020; Janßen and Wan 2020; Drees and Sabourin 2021; Rohrbeck and Cooley 2021). These techniques are better-equipped to handle high-dimensional settings. This report comprises a critical review of such methods, with a focus on the framework for extremal PCA developed by Cooley and Thibaud (2019) and the algorithm for generating hazard event sets proposed by Rohrbeck and

[1] The code for this project is available at https://github.com/pawleymatthew/TFR.

Cooley (2021). We perform a detailed illustration of these methods using French rainfall data, identify their limitations, and outline directions for future research.

This report is organised as follows: Chapter 2 reviews the basic theory of extreme value theory within the framework of multivariate regular variation and overviews classical approaches for modelling the angular measure; Chapter 3 surveys statistical learning approaches that can handle high-dimensional settings, focussing on clustering and principal components analysis; Chapter 4 applies some of these methods to French rainfall data; Chapter 5 discusses their limitations and ideas for future research.

2
Background theory

2.1 Introduction

In a wide variety of applications, the study of extreme events is inherently a multivariate problem. Consider the context of climate extremes. We may be interested in studying the co-occurrence of extremes of several meteorological variables, e.g. wind speed and precipitation as in Vignotto et al. (2021). Alternatively, we might analyse spatial data concerning a single variable at several different locations, e.g. Bador et al. (2017) study temperature extremes in France. In each case the individual processes can be modelled using univariate techniques, but employing multivariate techniques that account for the inter-relationships between the processes' extremes will likely enhance the analysis. For example, in a spatial analysis we might expect that the data at one site could inform inferences at sites that are nearby or climatologically similar. Multivariate extreme value theory provides a rigorous mathematical framework for such analyses.

This chapter will present the basic mathematical theory of multivariate extremes within the framework of multivariate regular variation. Herein, we split the analysis of the tail of a random vector into two steps: modelling the marginals and modelling the extremal dependence structure. The former requires techniques from univariate extreme value theory, which is briefly summarised in Section 2.2. The phenomenon of extremal dependence - how do extremes in one component of a random vector relate to extremes in the other components? - is central to multivariate extremes. This concept, and the general theory of multivariate extremes, will be discussed in Section 2.3. We will find that the extremal dependence structure of a random vector is characterised by a measure, called the angular measure. Its estimation, particularly in high-dimensional settings, is inherently challenging and will be the focus of subsequent chapters.

2.2 Univariate extreme value theory

The theory of univariate extremes is well developed. The basic theory presented here is based on Beirlant et al. (2004) and Coles (2001); the reader is referred to these books for a more comprehensive overview.

Suppose we are interested in modelling the (upper) tail behaviour of a random variable X with distribution function F. Let $X_1, X_2, \ldots,$ be a sequence of independent observations of X. The key theoretical assumption underlying methods for modelling extremes is the so-called maximum domain of attraction condition (MDA): there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ and a non-degenerate random variable $Z \sim G$ such that

$$\frac{\max(X_1, \ldots, X_n) - b_n}{a_n} \xrightarrow{\mathcal{D}} Z, \qquad n \to \infty. \tag{2.1}$$

We say that X (or F) belongs to the maximum domain of attraction of Z (or G). The distribution G of Z belongs to a parametric family of distributions called the generalised extreme value (GEV) distribution (Fisher and Tippett 1928). Its distribution function takes the following parametric form

$$G(z) = \exp\left\{ - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\}, \tag{2.2}$$

defined on $\{z : (1 + \xi(z - \mu)/\sigma > 0)\}$, where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$ are location, scale and shape parameters, respectively. The GEV family encompasses three sub-families of distributions which exhibit qualitatively different tail behaviours. These sub-families are determined by the shape parameter: $\xi > 0$ gives the heavy-tailed Fréchet distribution, $\xi < 0$ corresponds to the negative Weibull distribution with a finite upper limit, and $\xi = 0$ (interpreted as the limit of (2.2) as $\xi \to 0$) corresponds to the exponential-tailed Gumbel distribution. Under the fundamental MDA assumption, there are two main strategies for modelling extremes: the block maxima method and the peaks-over-threshold method.

Models for block maxima are based on the representation (2.2). Given a series of independent observations $X_1, \ldots, X_n$, the data are divided into blocks of finite size m. It follows from the asymptotic theory that, for m sufficiently large, the block maxima are approximately GEV distributed.

A limitation of the block maxima approach is that it may fail to utilise some large observations, even though they may be informative for the tail. This motivates the alternative but intimately related peaks-over-threshold method, which considers the distribution of excesses over a given high threshold. If X is in the maximum domain of attraction of a GEV(μ, σ, ξ) distribution, then

$$\lim_{u \to \infty} \mathbb{P}\left(X > x + u \mid X > u\right) = \left(1 + \frac{\xi x}{\tilde{\sigma}}\right)_+^{-1/\xi}, \tag{2.3}$$

for $x > 0$, where $\widetilde{\sigma} = \sigma + \xi(u - \mu)$. The limiting conditional distribution is called the generalised Pareto distribution (GPD). Thus, for a sufficiently high fixed threshold u, exceedances by X of u are approximately GPD distributed. In practice, the threshold u is chosen using graphical diagnostic tools (Coles 2001, Section 4.3.1) or test-based approaches (Wadsworth 2016; Wadsworth and Tawn 2012).

2.3 Multivariate extreme value theory

2.3.1 Multivariate regular variation and the angular measure

We will study multivariate extremes within the framework of multivariate regular variation. Informally, a random vector is regularly varying if it is jointly heavy-tailed, meaning its joint tail decays according to a power law. A rigorous treatment of regular variation involves notions of convergence of measures; for details see Resnick (1987) and Resnick (2007). The regular variation framework is ubiquitous in multivariate extremes and in applications it is often assumed without validation, but a formal testing procedure has been developed (Einmahl et al. 2020). Although regular variation can be defined generally on $\mathbb{R}^d$, we restrict our attention to vectors on the non-negative orthant $\mathbb{R}_+^d = [0, \infty)^d$. This restriction implicitly assumes a directionality in the risk being assessed. Such directionality usually exists in applications. For example, an analysis of extreme rainfall is typically concerned with either heavy rainfall (flood risk) or scarce rainfall (drought risk), but not both.

The multivariate regular variation property implies that, for sufficiently large observations, the magnitude and direction of a random vector are approximately independent. Thus it is most simply described in terms of polar coordinates. Fix a norm $\|\cdot\|$ and denote the unit sphere on the non-negative orthant by $\mathbb{S}_+^{d-1} = \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\| = 1\}$. For any point $\boldsymbol{x} \in \mathbb{R}_+^d \setminus \{\boldsymbol{0}\}$, define the polar coordinate transformation $T(\boldsymbol{x}) = (\|\boldsymbol{x}\|, \boldsymbol{x}/\|\boldsymbol{x}\|) =: (r, \boldsymbol{w})$. The radial component r measures the distance of $\boldsymbol{x}$ from the origin and $\boldsymbol{w} \in \mathbb{S}_+^{d-1}$ is the associated angle. If a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}_+^d$ is regularly varying with tail index $\alpha > 0$, denoted $\boldsymbol{X} \in \mathrm{RV}_+^d(\alpha)$, then there exists a measure H on $\mathbb{S}_+^{d-1}$ such that for any Borel set $\mathcal{B} \subset \mathbb{S}_+^{d-1}$ and $z > 0$

$$\lim_{r \to \infty} \mathbb{P}\left(\|\boldsymbol{X}\| > rz, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \in \mathcal{B} \,\middle|\, \|X\| > r \right) = z^{-\alpha} H(\mathcal{B}). \tag{2.4}$$

At first glance, the requirement that $X_1, \ldots, X_d$ are heavy-tailed with a shared tail index may seem too restrictive to be useful. Indeed, this property is unlikely to be satisfied by the raw data, e.g. the GEV/GPD shape parameters of some components may be zero or negative, indicating light or finite tails. This issue is resolved by working with a transformed random vector obtained from the original random vector by performing suitable marginal transformations. Working with transformed marginals is common practice in extremes and can be theoretically justified (Resnick 1987, p. 265). A popular choice is Fréchet margins with shape parameter

$\xi = \alpha$, that is $\mathbb{P}(X_i \leq x) = \exp(-x^{-\alpha})$ for $x > 0$. This is achieved by transforming each X_i to $[-\log F_i(X_i)]^{-1/\alpha}$, where F_i is the marginal distribution function of X_i.

It follows from (2.4) that the extremal behaviour of a d-dimensional random vector $\boldsymbol{X}$ is fully characterised by two quantities: the (known) tail index α, which governs the heavy-tailedness, and the $(d-1)$-dimensional angular measure H, which contains all the information about the extremal dependence structure, i.e. the tail dependence between the components of $\boldsymbol{X}$. More details will be given in Section 2.3.2.

The radial-angular decomposition in (2.4) suggests - and provides a rigorous theoretical basis for - a practical strategy for extrapolating observed data to unobserved extreme levels: the angular components associated with observations above a high radial threshold may be used to estimate H. Unfortunately, this becomes inherently challenging in high dimensions ($d \gg 1$): it requires estimating a high dimensional measure using only those few extreme observations that contain an informative signal for the distributional tail. Methods for overcoming this difficulty are the primary focus of this project and subsequent chapters of this report.

2.3.2 Extremal dependence and summary measures

Extremal dependence is analogous to, but separate from, the notion of statistical dependence in non-extreme statistics. In particular, two random processes might appear independent in the standard sense but exhibit dependence in their extremes, e.g. daily price movements for two unrelated stocks that are susceptible to common shocks. In applications, the extremal dependence structure may be quite complex. For example, in a spatial analysis of climate extremes, it captures information such as the topography of the domain, the underlying physics of the climate system, and the distance between the spatial locations.

The concept of extremal dependence is formalised as follows. Define the tail correlation of X_i and X_j as

$$\chi_{ij} = \lim_{u\to 1} \chi_{ij}(u) = \lim_{u\to 1} \frac{\mathbb{P}(F_i(X_i) > u, F_j(X_j) > u)}{1-u} \tag{2.5}$$

where F_i denotes the distribution function of X_i. The variables X_i and X_j are said to be asymptotically independent if $\chi_{ij} = 0$ and asymptotically dependent if $\chi_{ij} > 0$, with $\chi_{ij} = 1$ corresponding to complete asymptotic dependence. In practice, the tail correlation is estimated by computing estimates $\hat{\chi}_{ij}(u)$ of $\chi_{ij}(u)$ for a range of quantiles u and selecting one as an approximation to χ_{ij} (Engelke and Ivanovs 2021, Figure 2). For any non-empty $I \subset \{1, \ldots, d\}$, we can define χ_I by extending (2.5) in the natural way. The subsets I with $\chi_I > 0$ correspond to groups of components that can be large simultaneously (with non-negligible probability).

The angular measure contains all information about the extremal dependence of $\boldsymbol{X}$. Asymptotic independence occurs if and only if H is concentrated on $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$, the vectors of the canonical basis of $\mathbb{R}^d$. It is important to note that asymptotic independence is a limiting case that cannot be attained in the multivariate regular variation paradigm. On the other hand, $\boldsymbol{X}$ exhibits complete asymptotic dependence if H places a single point mass at $\gamma\mathbf{1} \in \mathbb{S}_+^{d-1}$ for some normalising constant $\gamma > 0$ that depends on the choice of norm. Between these two degenerate cases, the angular measure (and corresponding dependence structure) can be very complicated. This motivates the use of simple summary measures.

The quantity χ defined in (2.5) is a popular summary measure for assessing the strength of tail dependence. A limitation is that it fails to discriminate between asymptotically independent distributions; a complementary measure $\bar{\chi}$ addresses this deficiency (Coles 2001, Section 8.4). The set of tail coefficients $\{\chi_{ij} : i, j = 1, \dots, d\}$ can be expressed in terms of H, but provides an incomplete description of the dependence structure of $\boldsymbol{X}$. In particular, χ only measures the strength of pairwise dependencies.

An alternative summary measure is the extremal dependence measure (EDM), defined in the bivariate case by Larsson and Resnick (2012). For a regularly varying random vector $\boldsymbol{X}$ with angular measure H on $\mathbb{S}_+^{d-1}$, the EDM between X_i and X_j is given by

$$\mathrm{EDM}(X_i, X_j) = \int_{\mathbb{S}_+^{d-1}} w_i w_j \, \mathrm{d}H(\boldsymbol{w}). \tag{2.6}$$

It is easy to see that $\mathrm{EDM}(X_i, X_j) = 0$ in the case of asymptotic independence and that $\mathrm{EDM}(X_i, X_j)$ attains its maximum in the case of complete asymptotic dependence. Again, the set $\{\mathrm{EDM}(X_i, X_j) : i, j = 1, \dots, d\}$ is fully determined by the angular measure, but only contains summary information about the pairwise dependencies.

2.3.3 Classical models

The family of possible extremal dependence structures is in one-to-one correspondence with the class of angular measures, i.e. the class of positive measures on the unit simplex satisfying a number of mean constraints (Beirlant et al. 2004, Section 8.2.3). Unfortunately, this class is very large and does not admit a finite-dimensional parametrisation. A popular approach is to perform inference within a well-chosen parametric sub-model, constructed in such a way that such that the parametric sub-family generates a wide class of (valid) dependence structures. However, the trade-off between model flexibility and model parsimony is difficult to manage. The logistic model, proposed initially in the bivariate setting by Gumbel (1960), has only one parameter and is symmetric in all components. This is too restrictive to capture the complex dependence structure encountered in many applications. An asymmetric extension of the logistic model, developed by Tawn (1988), is more flexible but at the expense of the dependence structure being defined by 2^d parameters. The Hüsler-Reiss (HR) distribution (Hüsler and Reiss 1989) is parametrised by a conditionally negative definite matrix $\Gamma \in \mathbb{R}^{d \times d}$, called the variogram, whose entries Γ_{ij} control the strength of extremal dependence between pairs of components of $\boldsymbol{X}$. However, the number of parameters is $d(d+1)/2$, which grows quickly as d increases. A more comprehensive overview of parametric models for multivariate extremes can be found in Gudendorf and Segers (2010). Generally, these models are ill-suited to high dimensions because they are either too inflexible or the number of parameters increases too rapidly. The curse of dimensionality results in a lack of parsimony and impedes inference. The latter issue is exacerbated by the fact that we have only a limited number of extreme observations at our disposal. Parsimonious models have been developed specifically for spatial applications, but they require prior domain knowledge and stationarity assumptions (Wackernagel 1995). This is too restrictive and precludes a purely data-driven approach.

More recently, semi-parametric models have been explored. Boldi and Davison (2007) propose a constrained mixture of Dirichlet distributions, where the number of mixture components k is allowed to vary. The number of parameters is of order kd. Hanson et al. (2017) generalise the model to allow H-mass to be placed at the boundaries of $\mathbb{S}_+^{d-1}$, and another extension by de Carvalho and Davison (2014) permits a mixture of different spectral distributions (not just Dirichlet). The drawbacks of these approaches are primarily practical in nature: model fitting is typically performed using a reversible jump MCMC (RJMCMC) algorithm that is cumbersome to implement and the computations become infeasible for large k.

The limitations of these classical models motivate research into alternative approaches using techniques from statistical learning. These will be reviewed in next chapter.

3

Literature review: extremal dependence in high dimensions

3.1 Introduction

The tail dependence of a d-dimensional random vector is characterised by a $(d-1)$-dimensional angular measure, whose estimation is challenging, especially in high dimensions. This difficulty has given rise to a new area of research concerning statistical learning methods for analysing extremal dependence. In some cases, these approaches are limited to a low or moderate number of dimensions (Goix et al. 2017; Simpson et al. 2019; Meyer and Wintenberger 2021). A more promising line of research focusses on adapting popular unsupervised learning techniques, such as clustering and principal component analysis (PCA), for extremes (Bernard et al. 2013; Chautru 2015; Cooley and Thibaud 2019; Drees and Sabourin 2021; Fomichov and Ivanovs 2020; Janßen and Wan 2020; Rohrbeck and Cooley 2021). If the angular measure has a sparse structure, then such methods can facilitate dimension reduction.

3.1.1 Notions of sparsity

Clustering and principal components analysis are popular tools in multivariate statistics used to detect low-dimensional structures in data (James et al. 2021). Their application implicitly assumes a notion of sparsity, so that the object of interest can be expressed as (or at least well-approximated by) a lower-dimensional object. PCA assumes that certain linear combinations of the variables tend to be more likely than others, so that the data can be projected onto a lower-dimensional subspace while incurring minimal information loss. Clustering assumes that the observations can be partitioned into distinct, homogeneous subgroups. Thus, in order to adapt/apply these techniques to the analysis of multivariate tails, we require that the angular measure exhibits some sparse structure. Specifically, we assume that the dimension of the support of the angular measure H is much less than d. In many applications, only a small number ($\ll 2^d - 1$) of subsets of components are likely to be simultaneously large, so this assumption is usually reasonable. For example, heavy rainfall events tend to be localised so that only groups of neighbouring sites are jointly impacted. Notions of sparsity are discussed in more detail in Engelke and Ivanovs (2021).

3.1.2 Data simulation

Some of the methods in this chapter will be illustrated using synthetic data generated from a d-variate max-stable Hüsler-Reiss (HR) distribution. Full details of the simulation framework and methodology are given in Appendix A. We generate $n = 5000$ samples in $d = 75$ dimensions with $\beta = 2$. The dependence structure is constructed so that the angular measure is supported on fives faces of $\mathbb{S}_+^{d-1}$, whose dimensions are 20, 20, 20, 10 and 5. Components belonging to different groups are asymptotically independent. Components belonging to the same group exhibit asymptotic dependence with varying strengths. Figure 3.1 shows the entries of the matrix of tail correlation coefficients χ_{ij}. The darker coloured cells corresponds to pairs of components with strong extremal dependence. The zero entries in the off-diagonal blocks is due to asymptotic independence between components in different subgroups. Figure 3.2 shows some exploratory plots, produced by projecting the sample angles of large observations onto various faces of the unit sphere $\mathbb{S}_+^2 = \{\boldsymbol{x} \in \mathbb{R}_+^3 : \|\boldsymbol{x}\|_2 = 1\}$. In the left plot, all components are dependent, so we observe a number of points in the middle of the interior of the simplex. In the middle plot, only two of the components are asymptotically dependent. Extremes in X_1, and X_2 are dependent and therefore likely to co-occur, but they are independent of extremes X_{21}. As a result, the points are concentrated along the bottom edge and upper corner. In the third plot, all components are asymptotically independent and extremes tend to occur individually.

3.2 Clustering

Clustering refers to the task of partitioning a set of objects into distinct, homogeneous subgroups, called clusters (James et al. 2021). Clustering is an unsupervised learning problem: the goal is to discover structure from a set of observations. Fundamental to any clustering method is the notion of (dis)similarity: to speak of homogeneous subgroups and heterogeneous observations we must define what it means for two objects to be similar/different. This is a domain-specific consideration. Techniques for cluster analysis in a non-extreme setting, such as k-means and hierarchical clustering, are well established and have been applied in many fields. Recent work adapts these methods for extremes to facilitate exploration of the extremal dependence structure.

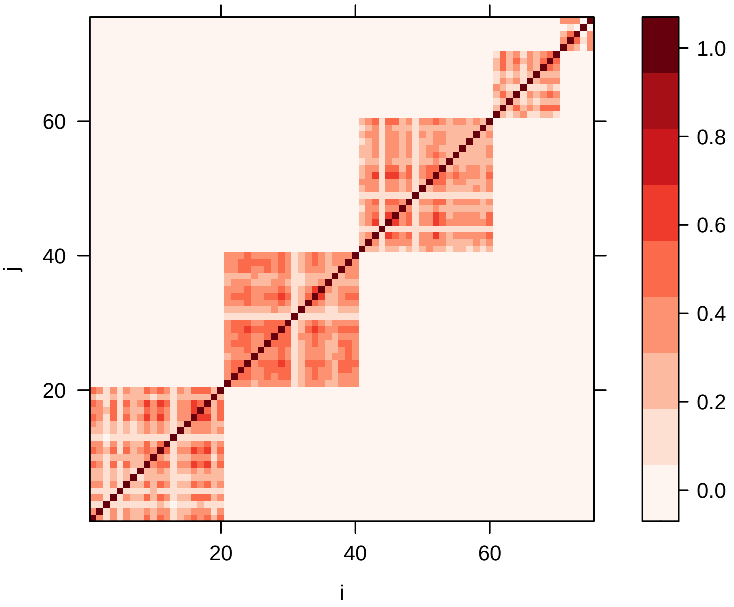

Figure 3.1: The matrix of true pairwise tail correlation coefficients χ_{ij} for the simulated data.

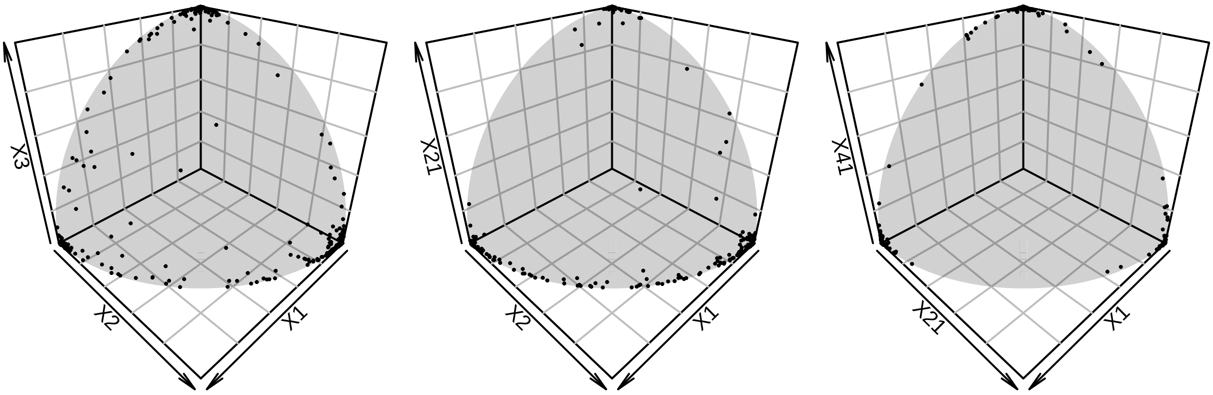

Figure 3.2: Projections of the sample angles corresponding to largest 5% of the simulated observations onto various faces of $\mathbb{S}_+^2$.

3.2.1 Clustering based on pairwise extremal dependence measures

The first class of methods clusters components of $\boldsymbol{X}$ according to some dissimilarity defined in terms of a measure of pairwise extremal dependence. This approach was initially proposed by Bernard et al. (2013). Given a sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, they define the pairwise dissimilarity between components X_i and X_j as

$$d_{ij} = \frac{1 - \chi_{ij}}{2(3 - \chi_{ij})} \tag{3.1}$$

where χ_{ij} is the tail correlation coefficient defined in Section 2.3.2. This defines an interpretable distance, termed the F-madogram distance, ranging from $d_{ij} = 0$ in the case of complete asymptotic dependence to $d_{ij} = 1/6$ for asymptotic independence. The dissimilarity matrix

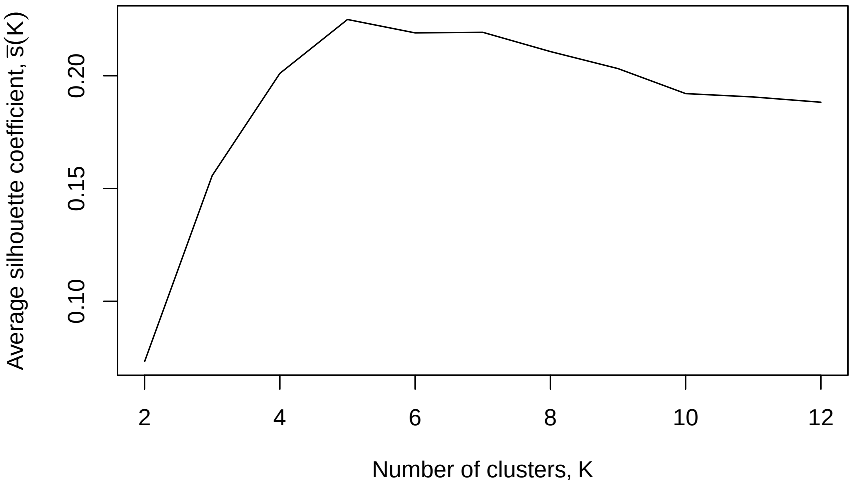

Figure 3.3: Silhouette analysis of the PAM clusters (based on F-madogram distances) for the simulated data. The average silhouette coefficient is maximised at $K = 5$.

$D = (d_{ij})$ can be estimated non-parametrically, and clustering is performed using the partitioning-around-medoids (PAM) algorithm of Kaufman and Rousseeuw (1990), which is closely related to k-means. The output is a set of $K \geq 2$ clusters such that tail dependence is stronger within groups than between groups. This method is very versatile: the dissimilarity measure in (3.1) can be tailored according to the particularities and aims of the study (Bador et al. 2017; Bracken et al. 2015; Mornet et al. 2017; Saunders et al. 2020; Vignotto et al. 2021). A limitation is that the number of clusters, K, is a hyperparameter that needs to be tuned. This is typically done by performing a silhouette analysis. The average silhouette coefficient, $\bar{s}(K)$, quantifies how informative a clustering is by comparing the intra- and inter-cluster distances (Rousseeuw 1987). The number of clusters can be chosen by comparing $\bar{s}(K)$ for a range of K values, as illustrated in Figure 3.3 for the simulated data. The plot indicates that there are five clusters, because $\bar{s}(K)$ is maximal at $K = 5$. This is in accordance with the known true dependence structure. However, the diagnostic is unlikely to be so conclusive in applications where the dependence structure is more complicated (see Bernard et al. 2013).

3.2.2 Bayesian hierarchical clustering

Another class of methods is based on hierarchical models, which are a natural approach for modelling spatial processes. Under this approach, the data is grouped into clusters at one or more levels. In a Bayesian hierarchical model, the number of groups, the cluster allocations, and the parameters for each cluster can be updated using Bayes' theorem. More details about hierarchical models can be found in Schervish (1995).

In the context of clustering for spatial extremes, several methods have been proposed. Carreau et al. (2017) propose a peaks-over-threshold model in which the GPD shape parameter is constant within clusters. The clustering facilitates information pooled across sites, reducing uncertainty in the estimation of ξ. However, their model gives no consideration to extremal dependence. Reich and Shaby (2019) allocate sites to K clusters using a spatial Potts model. The strength of spatial dependence over various scales is controlled by several parameters: K controls the limiting long-range dependence; the Potts parameter controls the rate of decay of dependence as a function of distance; a further parameter α controls the strength of dependence within the clusters. Neither Carreau et al. (2017) nor Reich and Shaby (2019) have a mechanism to update the number of clusters K, it must be chosen in advance. This deficiency is addressed by Rohrbeck and Tawn (2020). Given a number of clusters and a particular partition, their model is parametrised so that sites belonging to the same cluster have stronger extremal dependence, on average, and spatial dependence decays exponentially with distance, with a common rate of decay between clusters and a varying rate within clusters. The parameters, including K are updated by an RJMCMC algorithm.

3.2.3 Spherical clustering of extremal angles

Chautru (2015) propose exploring the angular measure by performing spherical clustering of the sample angles. Given sample data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$, let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n_{\text{exc}}} \in \mathbb{S}_+^{d-1}$ denote the set of angles (based on the Euclidean norm $\|\cdot\|_2$) associated with the n_{exc} observations above a given high radial threshold. The idea is to solve the optimisation problem

$$\min_{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_K \in \mathbb{S}_+^{d-1}} \sum_{i=1}^{n_{\text{exc}}} \min_{j=1,\ldots,K} d(\boldsymbol{w}_i, \boldsymbol{c}_j), \tag{3.2}$$

where K is a hyperparameter, $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K$ are the cluster centres (centroids), and $d : \mathbb{S}_+^{d-1} \times \mathbb{S}_+^{d-1} \to [0, \infty)$ measures the dissimilarity between two points on the unit sphere. Janßen and Wan (2020) take

$$d(\boldsymbol{u}, \boldsymbol{v}) = d_p(\boldsymbol{u}, \boldsymbol{v}) := 1 - (\boldsymbol{u}^T \boldsymbol{v})^p, \qquad (\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}_+^{d-1}), \tag{3.3}$$

with $p = 1$ and employ the spherical k-means clustering algorithm of Dhillon (2001) to locate the cluster centres. The case $p = 2$ is considered by Fomichov and Ivanovs (2020). Theoretical results and numerical experiments indicate that choosing $p = 2$ yields better results, especially when pairwise extremal dependence within clusters tends to be weak. Moreover, there are interesting links between spherical clustering with d_2 dissimilarity and extremal principal components analysis (the topic of the following section). For this reason, they refer to this special case of the general method as spherical k-principal-components clustering.

The centroids can be interpreted as the angular components of prototypical extreme events. A popular strategy for estimating the support of the angular measure is to assign positive H-mass to a face of the unit sphere if a centroid lies within a certain neighbourhood of that face (Chiapino and Sabourin 2017; Goix et al. 2017; Simpson et al. 2019; Chiapino, Sabourin and Segers 2018;

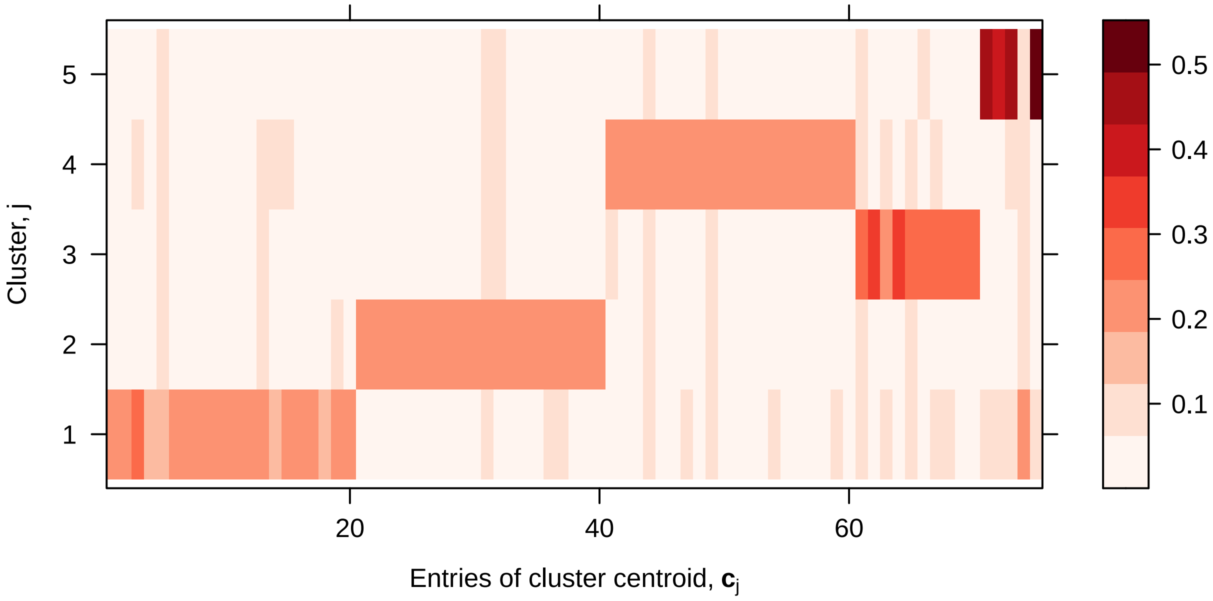

Figure 3.4: Representations of the cluster centroids obtained by applying spherical k-means clustering (with d_1 dissimilarity and $K = 5$) to the simulated data. The colours of the cells represent the size of the corresponding entry in the vector $\boldsymbol{c}_j$. Clustering is based on the sample angles corresponding to observations for which $r_i = \|\boldsymbol{x}_i\|_2$ exceeds the 95% quantile of $\{r_i : i = 1, \ldots, n\}$.

Meyer and Wintenberger 2021). Then the angular measure can be modelled by combining sub-models on each of these faces in a mixture model. However, these methods tend to be limited to moderate dimensions and they involve thresholding procedures that are highly sensitive to the choice of threshold. Figure 3.4 is a visual representation of the entries of the centroids obtained by applying the methodology of Janßen and Wan (2020) with $K = 5$ to the simulated data. On the whole, the centroids reflect the true dependence structure. However, the 74th component of $\boldsymbol{X}$ appears to be erroneously allocated to the first cluster; this is caused by the weak dependence between X_{74} and the other components in its 'true' subgroup (see Figure 3.1).

3.3 Principal components analysis

In the context of unsupervised learning, principal components analysis (PCA) refers to finding a low-dimensional linear subspace such that the data projected onto the subspace is as close as possible to the original data (James et al. 2021). In extremes, the goal is to find a low-dimensional subspace on which H is concentrated. Haug et al. (2009) were the first to do this, but they assume a specific parametric form for the extremal dependence structure. Drees and Sabourin (2021) and Cooley and Thibaud (2019) propose alternative statistical learning approaches that make no parametric assumption, in contrast.

3.3.1 PCA for the extremal angles

Drees and Sabourin (2021) consider the angles $\boldsymbol{W} = \boldsymbol{X}/\|\boldsymbol{X}\|$ and follow the standard PCA approach by projecting $\boldsymbol{W}$ onto low-dimensional linear subspaces $V \subset \mathbb{R}^d$. The optimal subspace is that which minimises the conditional risk

$$R_t(V) = \mathbb{E}\left[\|\Pi_V \boldsymbol{W} - \boldsymbol{W}\|_2^2 \middle| \|\boldsymbol{X}\| > t\right],$$

for some high threshold t, where $\Pi_V \boldsymbol{W}$ denotes the orthogonal projection of $\boldsymbol{W}$ onto V. The risk measures the reconstruction error incurred by reverting to a lower dimensional space. While the projection $\Pi_V \boldsymbol{W}$ does not necessarily lie on $\mathbb{S}_+^{d-1}$, this can be remedied by rescaling/shifting appropriately. In practice, given sample angles $\{\boldsymbol{w}_i = \boldsymbol{x}_i/\|\boldsymbol{x}_i\|\}_{i=1}^n$ the optimal subspace is estimated by considering the empirical risk

$$\hat{R}_t(V) = \frac{\sum_{i=1}^n \|\Pi_V \boldsymbol{w}_i - \boldsymbol{w}_i\|^2 \, \mathbb{1}(\|\boldsymbol{x}_i\| > t)}{\sum_{i=1}^n \mathbb{1}(\|\boldsymbol{x}_i\| > t)}. \tag{3.4}$$

As in standard PCA, there is a trade-off between reconstruction error and dimension reduction, because a high-dimensional subspace will yield better reconstructions than a lower-dimensional subspace. The number of dimensions is selected by comparing $R_t(\hat{V}_p)$ for a range of values $p \geq 1$, where $\hat{V}_p$ denotes the minimiser of $\hat{R}$ in the set of p-dimensional linear subspaces. The minimisers $\{\hat{V}_p : p \geq 1\}$ can be derived via a spectral analysis of the matrix of second mixed moments of $\boldsymbol{W}$. In high-dimensional simulation studies, Drees and Sabourin (2021) find that employing PCA on $\boldsymbol{W}$ does improve estimation of the angular measure (compared against a standard non-parametric estimator of H) but there are difficulties with choosing the number of dimensions.

3.3.2 PCA based on the tail pairwise dependence matrix

Roughly speaking, the theory of multivariate regular variation presented in Section 2.3.1 implies that performing PCA for $\boldsymbol{W}$ is essentially equivalent to performing PCA for $\boldsymbol{X}$ conditional on $\|\boldsymbol{X}\| > t$ (up to some rescaling and assuming standardised marginals) in the limit as $t \to \infty$. The latter interpretation underlies the approach originally developed by Cooley and Thibaud (2019) and later applied by Rohrbeck and Cooley (2021) to generate synthetic extreme events. The remainder of this report comprises a critical review of their methodologies: this section presents theoretical details; Chapter 4 illustrates their application; Chapter 5 discusses limitations and directions for future work.

Suppose $\widetilde{\boldsymbol{X}} \in \mathrm{RV}_+^d(2)$ is a random vector with Fréchet margins with shape $\xi = 2$, perhaps obtained by performing marginal transformations to the original random vector of interest $\boldsymbol{X}$. Let H_X denote the angular measure of $\widetilde{\boldsymbol{X}}$ on the L_2 unit sphere $\mathbb{S}_+^{d-1} = \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\|_2 = 1\}$. The tail dependence of $\widetilde{\boldsymbol{X}}$ may be summarised by the $d \times d$ matrix $\Sigma = (\Sigma_{ij})$ given by

$$\Sigma_{ij} = \int_{\mathbb{S}_+^{d-1}} w_i w_j \, \mathrm{d}H_X(\boldsymbol{w}). \tag{3.5}$$

The matrix Σ is called the tail pairwise dependence matrix (TPDM). It has been used to devise a suite of methods for analysing extremal dependence in various settings (Fix et al. 2021; Mhatre and Cooley 2021). The right-hand side of (3.5) is precisely the EDM from Section 2.3.2; the interpretation of the entries of Σ in terms of extremal dependence follows immediately. The choice of tail index $\alpha = 2$ endows the TPDM with properties analogous to that of the covariance matrix. In particular (Cooley and Thibaud 2019, Section 4):

1. Σ is non-negative definite.
2. The diagonal elements Σ_{ii} relate to the scale of the components of $\widetilde{\boldsymbol{X}}$. Specifically, for any $x > 0$,
$$\lim_{n\to\infty} n\mathbb{P}\left(\frac{\widetilde{X}_i}{\sqrt{n}} > x\right) = \frac{\Sigma_{ii}}{x^2}.$$
3. The sum of the diagonal elements of Σ equals the total mass of H_X.
4. Two components $\widetilde{X}_i$ and $\widetilde{X}_j$ are asymptotically independent if and only if $\Sigma_{ij} = 0$.

Following the approach of standard PCA, we consider the eigendecomposition $\Sigma = UDU^T$, where D is a diagonal matrix of real eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ and $U \in \mathbb{R}^{d\times d}$ is a unitary matrix whose columns are the corresponding eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$. Let $\tau : \mathbb{R} \to [0, \infty)$ denote the softplus function defined by $\tau(x) = \log(1 + \exp(x))$ for $x \in \mathbb{R}$. This function allows us to map between $\mathbb{R}^d$ and $\mathbb{R}^d_+$ without interfering with the tails. Applying the transform component-wise, the vectors $\tau(\boldsymbol{u}_1), \ldots, \tau(\boldsymbol{u}_d)$ form an orthonormal basis for $\mathbb{R}^d_+$. Moreover, the basis is ordered such that each vector corresponds to the direction of maximum scale after accounting for the information contained in the previous basis vectors. The eigenvalues measure the amount of scale explained by these directions.

The extremal principal components of $\widetilde{\boldsymbol{X}}$ are defined by

$$\boldsymbol{V} = \boldsymbol{U}^T \tau^{-1}(\widetilde{\boldsymbol{X}}). \tag{3.6}$$

Unlike $\widetilde{\boldsymbol{X}}$, the extremal principal components lie in the entire space $\mathbb{R}^d$. By reversing (3.6), it can be shown that

$$\widetilde{\boldsymbol{X}} = \tau\left(\sum_{i=1}^{d} V_i \boldsymbol{u}_i\right). \tag{3.7}$$

Truncating the sum in (3.7) yields the optimal (in terms of L_2-distance) low-dimensional projections of $\widetilde{\boldsymbol{X}}$ (Engelke and Ivanovs 2021).

The random variable $\boldsymbol{V}$ is multivariate regularly varying with tail index $\alpha = 2$ (Cooley and Thibaud 2019, Lemma A4). Furthermore, the random variable $\|\boldsymbol{V}\|_2$ follows a Fréchet distribution with $\mathbb{P}(\|\boldsymbol{V}\|_2 \leq t) = \exp[-(t/d)^{-2}]$ for $t > 0$, due to the norm preservation property of the unitary matrix U used in the projection in (3.6). Let H_V, defined on the entire unit sphere $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$, denote the angular measure of $\boldsymbol{V}$, and define the TPDM $\widetilde{\Sigma}$ of $\boldsymbol{V}$ in the natural way, i.e. by replacing H_X with H_V in (3.5). By Proposition 6 in Cooley and Thibaud (2019), we have $\widetilde{\Sigma}_{ii} = \lambda_i$ for $i = 1, \dots, d$ and $\widetilde{\Sigma}_{ij} = 0$ for $i \neq j$. Unfortunately, now $\widetilde{\Sigma}_{ij} = 0$ does not imply asymptotic independence between V_i and V_j.

Rohrbeck and Cooley (2021) show how this framework can be exploited to generate samples from the tail of $\widetilde{\boldsymbol{X}}$ (and ultimately $\boldsymbol{X}$, by applying the inverse of the initial marginal transformations). The key idea is to sample from the tail distribution of $\boldsymbol{V}$ instead of directly sampling from the tail of $\widetilde{\boldsymbol{X}}$, which would require estimating H_X. Unfortunately, since the extremal principal components are asymptotically dependent, sampling from the tail distribution of $\boldsymbol{V}$ still requires the estimation of the $(d-1)$-dimensional measure H_V. However, the crucial difference is that the components of $\boldsymbol{V}$ are ordered so that its first few components contain the most information about the structure of extreme events. In particular, for some high threshold r_V, the distribution of

$$\boldsymbol{W} = \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|_2} \,\bigg|\, \|\boldsymbol{V}\|_2 > r_V$$

is estimated by combining a flexible model for the dependence structure of $(W_1, \dots, W_m)$ and a restrictive model for the dependence structure of $(W_{m+1}, \dots, W_d)$. The value $1 \leq m \leq d$ is selected according to the number of eigenpairs needed to capture the large-scale extremal behaviour of $\widetilde{\boldsymbol{X}}$. Provided the angular measure of $\widetilde{\boldsymbol{X}}$ exhibits a sparse structure, $m \ll d$ should be sufficient, so we reap the benefits of dimension reduction. Asymptotic dependence between the principal components dictates that, for a realistic model, the large-scale behaviour $(W_1, \dots, W_m)$ and localised dynamics $(W_{m+1}, \dots, W_d)$ must be modelled jointly.

The first step is to estimate the distribution of a lower-dimensional random vector $\boldsymbol{Z} \in \mathbb{S}^{m+1}$ defined by $Z_i = W_i$ for $i = 1, \dots, m$ and

$$Z_{m+1} = \begin{cases} \sqrt{1 - \sum_{i=1}^m W_i^2}, & \text{if } W_{m+1} \geq 0 \\ -\sqrt{1 - \sum_{i=1}^m W_i^2}, & \text{if } W_{m+1} < 0 \end{cases}.$$

The first m components of $\boldsymbol{Z}$ describe the large-scale dependence structure, while the final component Z_{m+1} summarises local dynamics. The distribution of $\boldsymbol{Z}$ is then modelled by a suitable kernel density estimate for spherical data. Rohrbeck and Cooley (2021) choose a mixture of Mises-Fisher distributions (Hall et al. 1986): given observations $\{\boldsymbol{z}_i : i = 1, \dots, n_{\text{exc}}\}$ the estimated density is given by

$$\hat{h}(\boldsymbol{z}) = \frac{C(\kappa)}{n_{\text{exc}}} \sum_{i=1}^{n_{\text{exc}}} \exp(\kappa \boldsymbol{z}^T \boldsymbol{z}_i),$$

for $\boldsymbol{z} \in \mathbb{S}^{m+1}$, where $\kappa > 0$ is the bandwidth of the kernels and $C(\kappa)$ is a normalising constant.

Given a sample $\boldsymbol{z} \in \mathbb{S}^{m+1}$ from the fitted distribution for $\boldsymbol{Z}$, a sample $\boldsymbol{w} \in \mathbb{S}^{d-1}$ is derived by a nearest neighbours approach, by setting

$$\boldsymbol{w} = \left(z_1, \ldots, z_m, \left| \frac{z_{m+1}}{z^\star_{m+1}} \right| w_{m+1}, \ldots, \left| \frac{z_{m+1}}{z^\star_{m+1}} \right| w_d \right),$$

where $\boldsymbol{z}^\star$ is the nearest neighbour of $\boldsymbol{z}$ amongst $\{\boldsymbol{z}_i : i = 1, \ldots, n_{\text{exc}}\}$. The simulated principal components $\boldsymbol{v} \in \mathbb{R}^d$ are given by $\boldsymbol{v} = r\boldsymbol{w}$, where $r = \|\boldsymbol{V}\|_2$ is sampled from a Fréchet distribution with $\mathbb{P}(\|\boldsymbol{V}\|_2 \leq t) = \exp[-(t/d)^{-2}]$. Finally, a sample $\widetilde{\boldsymbol{x}} \in \mathbb{R}^d_+$ from the approximate tail distribution of $\widetilde{\boldsymbol{x}}$ is obtained by applying the inverse of (3.6) to $\boldsymbol{v}$.

The next chapter will illustrate how these methods are applied and provide more practical details regarding their implementation.

4

Application to the France rainfall data

This chapter illustrates how extremal PCA is applied using a real-world dataset from climatology. In Section 4.2, we perform an analysis of extreme rainfall in France using the framework of Cooley and Thibaud (2019). Then the sampling algorithm of Rohrbeck and Cooley (2021) is applied in Section 4.3 to generate synthetic extreme rainfall events. The underlying theoretical details for these methods were summarised in Section 3.3.2; the emphasis of this chapter is showing how these tools are used in practice. From this exercise we can identify limitations/deficiencies that could be addressed by future research, which will be discussed in the next chapter.

4.1 Data description

The French precipitation dataset consists of the weekly maxima of hourly precipitation measured at $d = 92$ weather stations in France during the autumn season (September-November) between 1993 and 2011[1]. The weather stations provide a fairly complete and homogeneous coverage of France, as illustrated in Figure 4.1. There are $T = 228$ weeks of recorded rainfall maxima at each station, with no missing data. Figure 4.2 shows the time series recorded at Montereau-sur-le-Jard (a commune near Paris) and Tarbes–Lourdes–Pyrénées Airport (on the southwest coast). These stations are marked in red in Figure 4.1. An important assumption underlying our methods is that there is no serial dependence within the time series. This assumption seems reasonable for hourly precipitation maxima taken over weekly periods and the autocorrelograms in Figure 4.2 provide further assurance.

[1] The data was collected by Météo-France, the French meteorological service, and is available from the homepage of the second author of Bernard et al. (2013) at https://www.lsce.ipsl.fr/Phocea/Pisp/visu.php?id=109&uid=naveau.

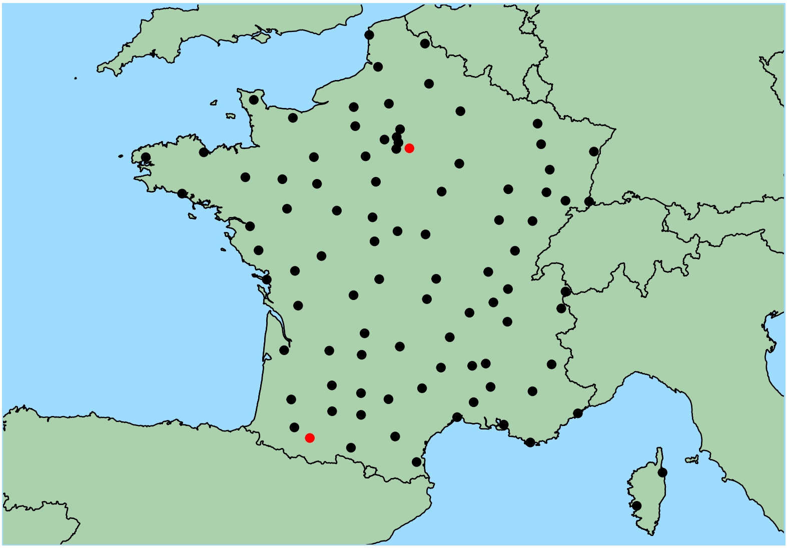

Figure 4.1: The locations of the 92 Météo-France weather stations. Montereau-sur-le-Jard and Tarbes–Lourdes–Pyrénées Airport are marked in red.

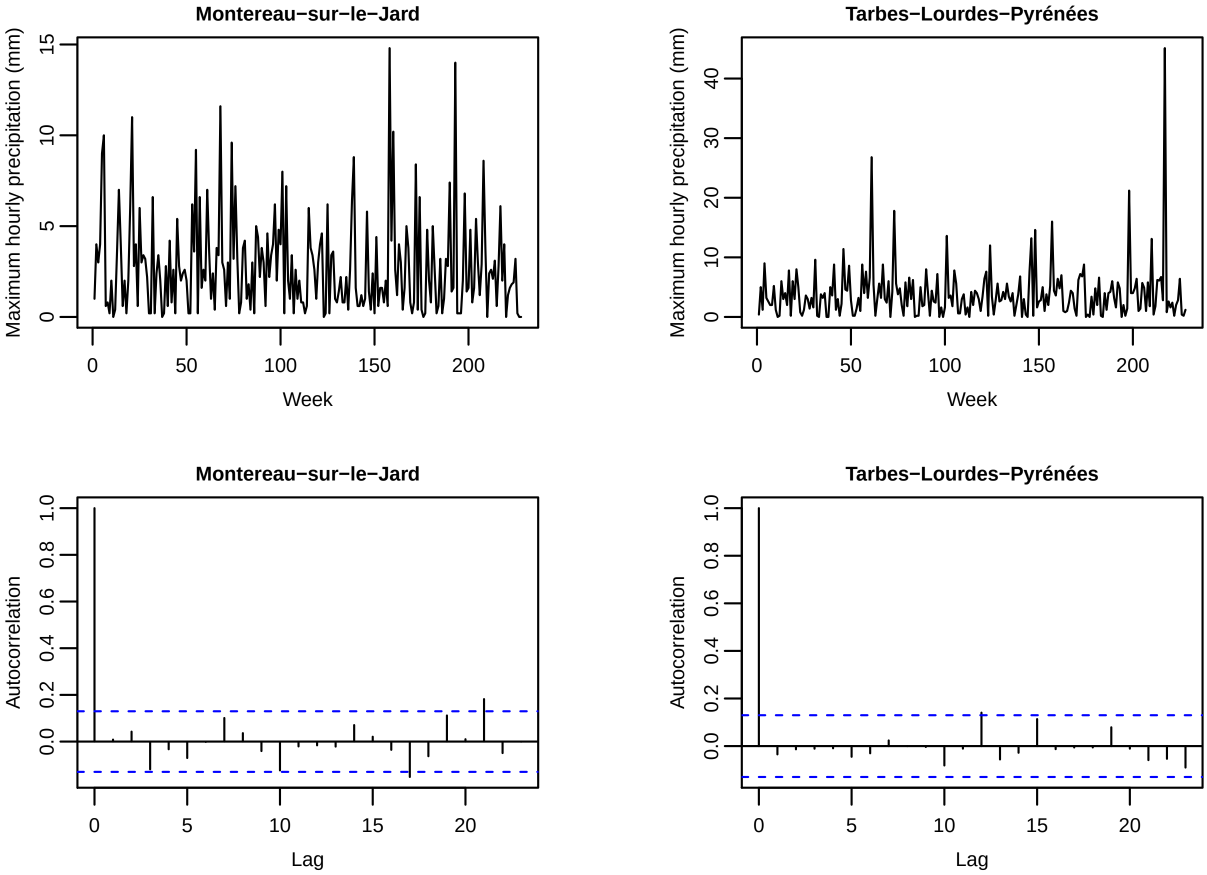

Figure 4.2: Top: weekly maxima of hourly precipitation recorded at Montereau-sur-le-Jard (left) and Tarbes–Lourdes–Pyrénées Airport (right). Bottom: the autocorrelograms for each time series. The dashed lines indicate the 95% confidence intervals.

4.2 Analysing extremal dependence using extremal PCA

4.2.1 Data preprocessing

Let $X_{t,i}$ denote the random variable representing the maximum hourly precipitation at station i during week t, for $i = 1, \ldots, d$ and $t = 1, \ldots, T$, and let $\boldsymbol{X}_t = (X_{t,1}, ..., X_{t,d})$. We apply a transformation to $\boldsymbol{X}_t$ in order to obtain a random variable $\widetilde{\boldsymbol{X}}_t = (\widetilde{X}_{t,1}, \ldots, \widetilde{X}_{t,d})$ that is regularly varying with tail index $\alpha = 2$ and has Fréchet margins, $\mathbb{P}\left(\widetilde{X}_{t,i} \leq x\right) = \exp(-x^{-2})$ for $x > 0$. This is achieved by defining

$$\widetilde{X}_{t,i} = \left[-\log \hat{F}_i(X_{t,i})\right]^{-1/2} \tag{4.1}$$

for $i = 1, \ldots, d$ and $t = 1, \ldots, T$, where $\hat{F}_i$ is an estimate of the distribution function for the weekly precipitation maxima at station i. Here we take $\hat{F}_i$ to be the empirical CDF obtained by a rank transform; a more sophisticated semi-parametric approach is outlined in Rohrbeck and Cooley (2021). Let $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_T$ be the set of transformed observations obtained by applying the transformation (4.1) to the original measurements $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$.

4.2.2 Estimation of the TPDM

Next, we estimate the tail pairwise dependence matrix (TPDM), Σ. There are two estimators used in the literature. The first approach, used by Cooley and Thibaud (2019) and Rohrbeck and Cooley (2021), is to threshold observations based on the entire vector. Define $r_t = \|\widetilde{\boldsymbol{x}}_t\|_2$ and set $r^\star$ as some high quantile of $\{r_t : t = 1, \ldots, T\}$. Let $\mathcal{T}^\star = \{t : r_t > r^\star\}$ represent the set of times at which extreme events occurred and denote by $\boldsymbol{w}_t = \widetilde{\boldsymbol{x}}_t / r_t$ the associated angular components. Then the vector-based TPDM estimate is defined as the $d \times d$ matrix $\hat{\Sigma}^{(v)}$ with entries

$$\hat{\Sigma}^{(v)}_{i,j} = \frac{d}{|\mathcal{T}^\star|} \sum_{t \in \mathcal{T}^\star} w_{t,i} w_{t,j}. \tag{4.2}$$

The second approach, adopted by Jiang et al. (2020), is based on pairwise radial thresholds. For $i, j \in \{1, \ldots, d\}$, define the radius $r_{t,i,j} = \|(\widetilde{x}_{t,i}, \widetilde{x}_{t,j})\|_2$. Choose a high radial threshold $r^\star_{i,j}$ of $\{r_{t,i,j} : t = 1, \ldots, T\}$ and define $\mathcal{T}^\star_{i,j} = \{t : r_{t,i,j} > r^\star_{i,j}\}$. In a slight abuse of notation, define the angular components $(w_{t,i}, w_{t,j}) = (x_{t,i}, x_{t,j})/r_{t,i,j}$. Then the pairs-based TPDM estimate is defined as the $d \times d$ matrix $\hat{\Sigma}^{(p)}$ with entries

$$\hat{\Sigma}^{(p)}_{i,j} = \frac{2}{|\mathcal{T}^\star_{i,j}|} \sum_{t \in \mathcal{T}^\star_{i,j}} w_{t,i} w_{t,j}. \tag{4.3}$$

This matrix is not guaranteed to be positive definite, which is required for PCA. This issue is resolved by taking the final estimate $\hat{\Sigma}^{(p)}$ as the nearest (in terms of Frobenius norm) positive definite matrix to the matrix obtained from (4.3). Another issue with this estimator is that

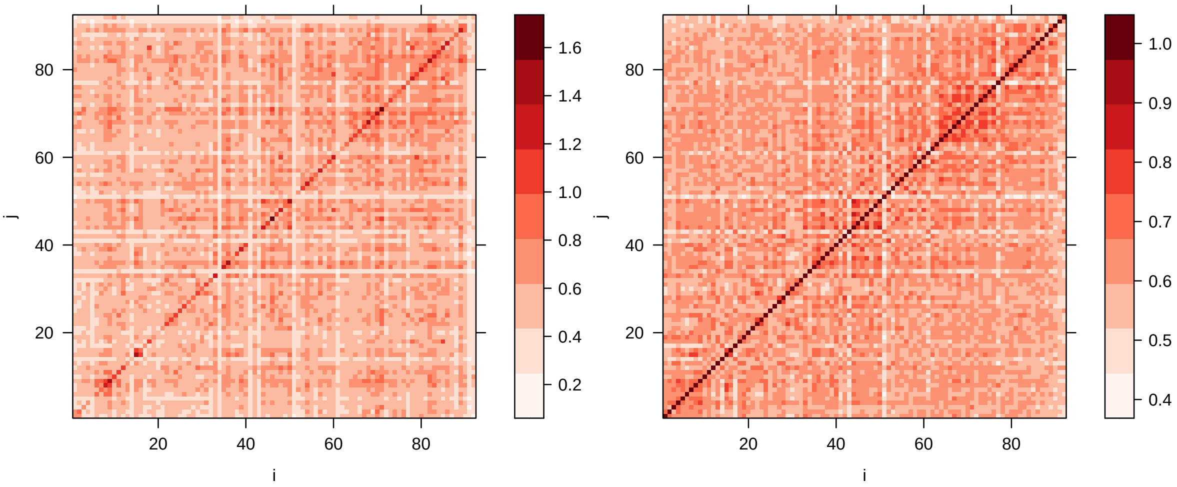

Figure 4.3: Entries $\hat{\Sigma}_{ij}$ of the vector-based (left) and pairs-based (right) TPDM estimates. The radial thresholds were set as the empirical 85% quantile.

the interpretation of the diagonal elements in terms of the components' scales is lost. In fact, it is easy to show that $\hat{\Sigma}^{(p)}_{i,i} = 1$ for all i.

The key difference between the estimators is that $\hat{\Sigma}^{(v)}_{i,j}$ uses observations that are extreme globally (even if X_i and X_j themselves are not large) while $\hat{\Sigma}^{(p)}_{i,j}$ is based on events for which X_i and X_j are large (irrespective of the size of the other components). For this reason, the pairs-based estimator might be preferred for analyses where behaviour is expected to be highly localised, e.g. if the study region is large. Our spatial domain is larger than those of Cooley and Thibaud (2019) and Rohrbeck and Cooley (2021) but smaller than that of Jiang et al. (2020), so it is not obvious which estimator we should prefer. Both estimates for Σ, based on radial thresholds taken at the empirical 85% quantile, are illustrated in Figure 4.3. The stations are ordered by increasing longitude (i.e. from south to north). Apart from the differences in scaling the two matrices are quite similar, structurally speaking. Subsequent examination of their respective eigenpairs revealed no discernible differences. Therefore we proceed with the vector-based estimate on the basis that it is simpler, faster to compute, and satisfies all the properties of a TPDM. Henceforth, $\hat{\Sigma}$ denotes the vector-based estimate, unless stated otherwise.

Huser et al. (2016) note that threshold-based estimators have a tendency to overestimate dependence when the true dependence is weak/moderate. In order to mitigate this bias, Fix et al. (2021) modify the estimator such that $\hat{\Sigma}_{ij}$ is close to zero if the distance between stations i and j is large. This assumption is reasonable due to the localised nature of extreme rainfall events. For simplicity, we opt not incorporate any bias correction into our estimate, but it is worth investigating in future.

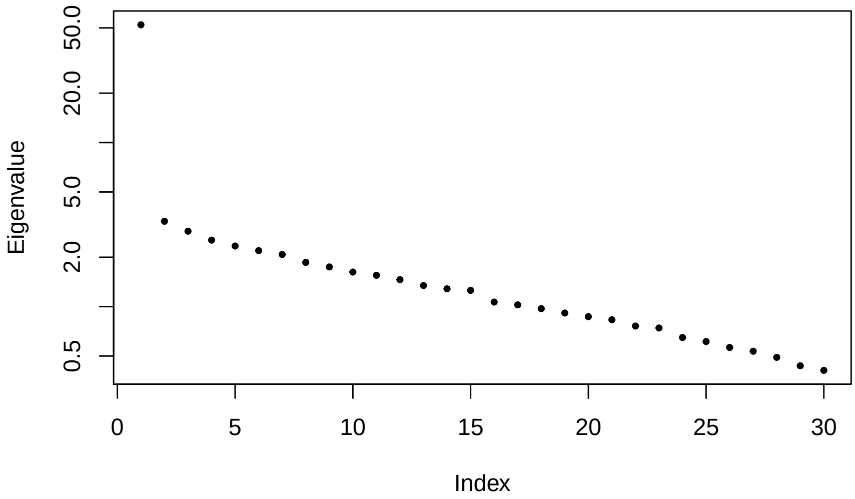

Figure 4.4: Eigenvalues of the first 30 eigenvalues of the TPDM on a log scale.

4.2.3 Examining the eigenvalues

We perform an eigendecomposition of $\hat{\Sigma}$ to obtain $\hat{\Sigma} = \hat{U}\hat{D}\hat{U}^T$, where $\hat{D}$ is a diagonal matrix of real eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$ and $\hat{U}$ is a $d \times d$ unitary matrix whose columns are the corresponding eigenvectors $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_d \in \mathbb{R}^d$.

Figure 4.4 shows a scree plot of the first 30 eigenvalues of $\hat{\Sigma}$. The first eigenvalue is very large; thereafter the values slowly decrease to zero. The first six eigenvalues are $(\lambda_1, \ldots, \lambda_6) = (52.2, 3.3, 2.9, 2.5, 2.3, 2.2)$. This implies that the first six principal components account for 71% of the total scale of $\widetilde{\boldsymbol{X}}$. For comparison, studies of extreme rainfall in the continental US and a small region of the UK found that the first six eigenvectors accounted for 41% and 88% of the scale, respectively (Jiang et al. 2020; Rohrbeck and Cooley 2021). The differences between these values are in accordance with the sizes of the study regions and the degree to which extreme events are localised. We conclude that extreme precipitation in France is quite localised and a large number of eigenvectors will be required to reconstruct small-scale features of extreme events.

4.2.4 Interpreting the eigenvectors

The leading eigenvectors reveal the large-scale spatial behaviour of extreme rainfall events in France. Spatial representations of first six eigenvectors are shown in Figure 4.5. The first eigenvector, $\hat{\boldsymbol{u}}_1$, is positive and accounts for the magnitude of extreme events. The spatial patterns in $\hat{\boldsymbol{u}}_1$ do not necessarily reflect the marginal law behaviour at each site, cf. Figure 1b in Bernard et al. (2013). This is primarily because we are working with standardised data, for which ‘extreme’ should be interpreted as ‘extreme relative to the climate at that location’. The second eigenvector reveals a north-south signal. This divide can be justified climatologically: extreme events in the south are due to thunderstorms caused by warm air interacting with the mountainous regions (Pyrénées/Cévennes/Alps); heavy rainfall in the north is produced

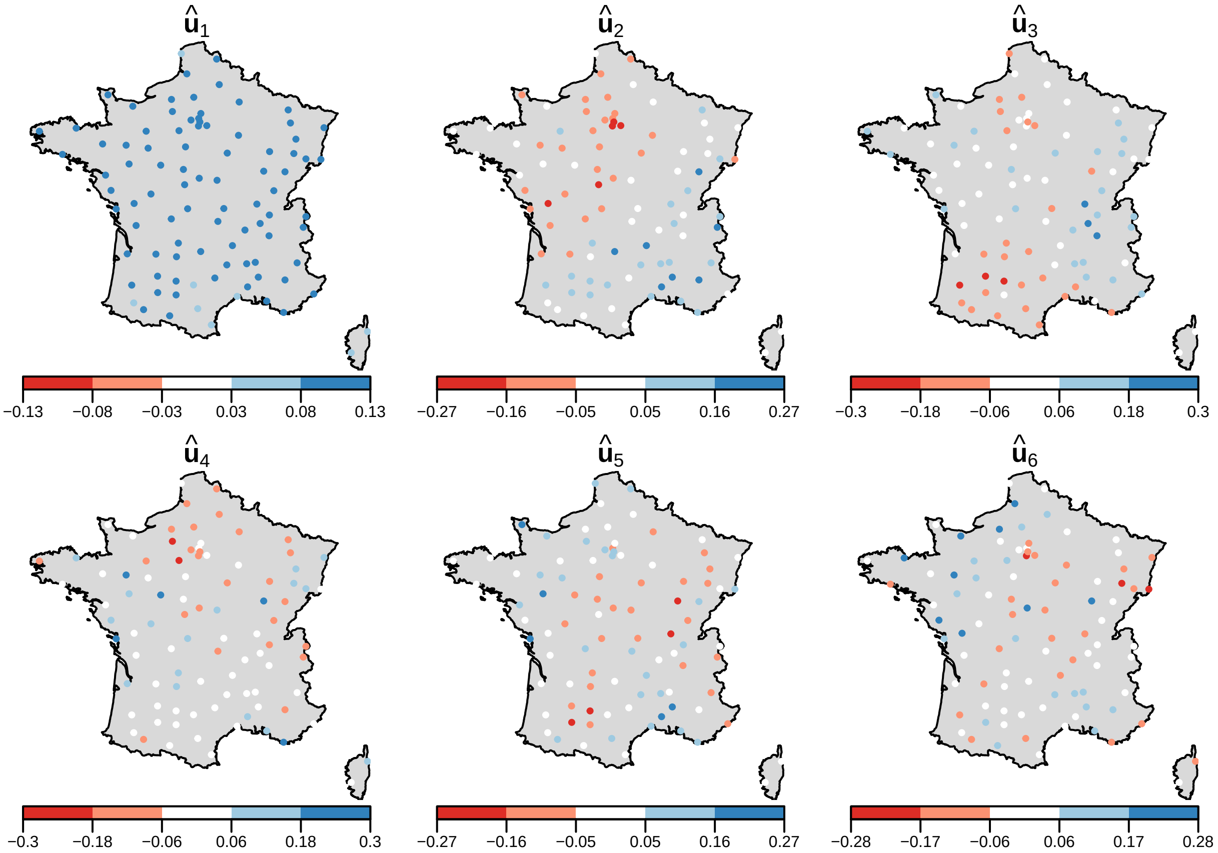

Figure 4.5: First six eigenvectors of the TPDM.

by midlatitude perturbations (Bernard et al. 2013). The third eigenvector shows a strong negative signal on the south-west coast. An east-west divide along the southern coast makes sense because extreme events originating in Toulon or Nice tend to not to affect areas to the west of Montpellier. The subsequent eigenvectors reveal more localised patterns in extremal behaviour and are more difficult to interpret.

4.2.5 Analysing the extremal principal components

Finally, we study the extremal principal components given by $\boldsymbol{v}_t = \hat{\boldsymbol{U}}^T\tau^{-1}(\widetilde{\boldsymbol{x}}_t)$ for $t = 1, \ldots, T$. The elements of $\boldsymbol{v}_t \in \mathbb{R}^d$ are the stochastic basis coefficients when $\widetilde{\boldsymbol{x}}_t$ is decomposed into the basis $\tau(\hat{\boldsymbol{u}}_1), \ldots, \tau(\hat{\boldsymbol{u}}_d)$. Figure 4.6 shows time series plots of the extremal principal components associated with the first three eigenvectors. The points highlighted in red correspond to the weeks for which $r_t = \|\widetilde{\boldsymbol{x}}_t\|_2$ exceeds the 95% quantile of $\{r_t : t = 1, \ldots, T\}$. Importantly, note that the extremes of $\boldsymbol{V}$ tend to coincide with the extremes of $\widetilde{\boldsymbol{X}}$.

The role of the principal components is most easily illustrated by exploring an observed extreme event. Figure 4.7 shows an array of plots associated with the event in week $t = 181$. The top left plot is a spatial representation of the event (after marginal transformation). Extreme rainfall intensities occurred in the mountainous region to the south/east of Lyon. The top

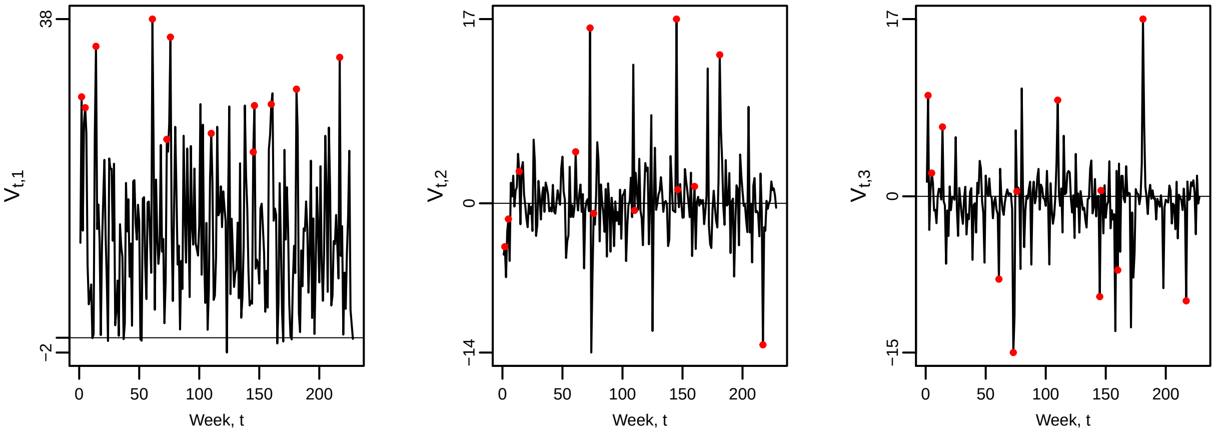

Figure 4.6: Time series of the observed extremal principal components. The weeks for which $r_t = \|\widetilde{\boldsymbol{x}}_t\|_2$ exceeds the 95% quantile of $\{r_t : t = 1, \ldots, T\}$ are highlighted in red.

right plot shows the elements of $\boldsymbol{v}_{181}$. Roughly speaking, the sizes of the $|V_{t,i}|$ tell us which eigenvectors are most important for capturing the event's spatial dynamics, since $V_{t,i}$ represents the coefficient associated with the ith eigenvector in the reconstruction. In this sense, the four most important eigenvectors are the first, second, third and seventeenth eigenvectors. Spatial representations of these eigenvectors are plotted in the second row of Figure 4.7 (the first eigenvector is omitted because it isn't particularly informative). The third eigenvector is hit by a large positive coefficient, which primarily allocates precipitation to the south east region. The negative coefficient $V_{181,2}$ diminishes the signal in the north and further boosts the signal in the south east. The seventeenth eigenvector captures very small-scale behaviours and serves to amplify the signal at the specific sites where the rainfall intensity was strongest (near Lyon and Avignon). The plots in the third and fourth rows of Figure 4.7 show a series of low-dimensional reconstructions of $\widetilde{\boldsymbol{x}}_{181}$ obtained by truncating the sum in (3.7). Note that the first four reconstructions have scaling issues: the intensities are generally too low because the omitted eigenvectors account for a non-negligible amount of scale. The eigenvalues of $\hat{\Sigma}$ decrease quite gradually, so this is only resolved once a large number of eigenvectors are added. The spatial attribution of rainfall improves as more principal components are added. The two-eigenvector reconstruction allocates rainfall too broadly; we know that the unused third eigenvector is important in restricting rainfall to the south east. The five eigenvector reconstruction looks much better but still overestimates rainfall in parts of the north. After 20 eigenvectors the reconstructed event looks quite accurate with only minor discrepancies, but the overall scale is still too low. The 45 eigenvector reconstruction matches the full basis reconstruction almost perfectly, because the omitted eigenvectors account for negligible scale ($\lambda_i \approx 0$ for $i > 45$) and their coefficients in the basis expansion are approximately zero (top right plot).

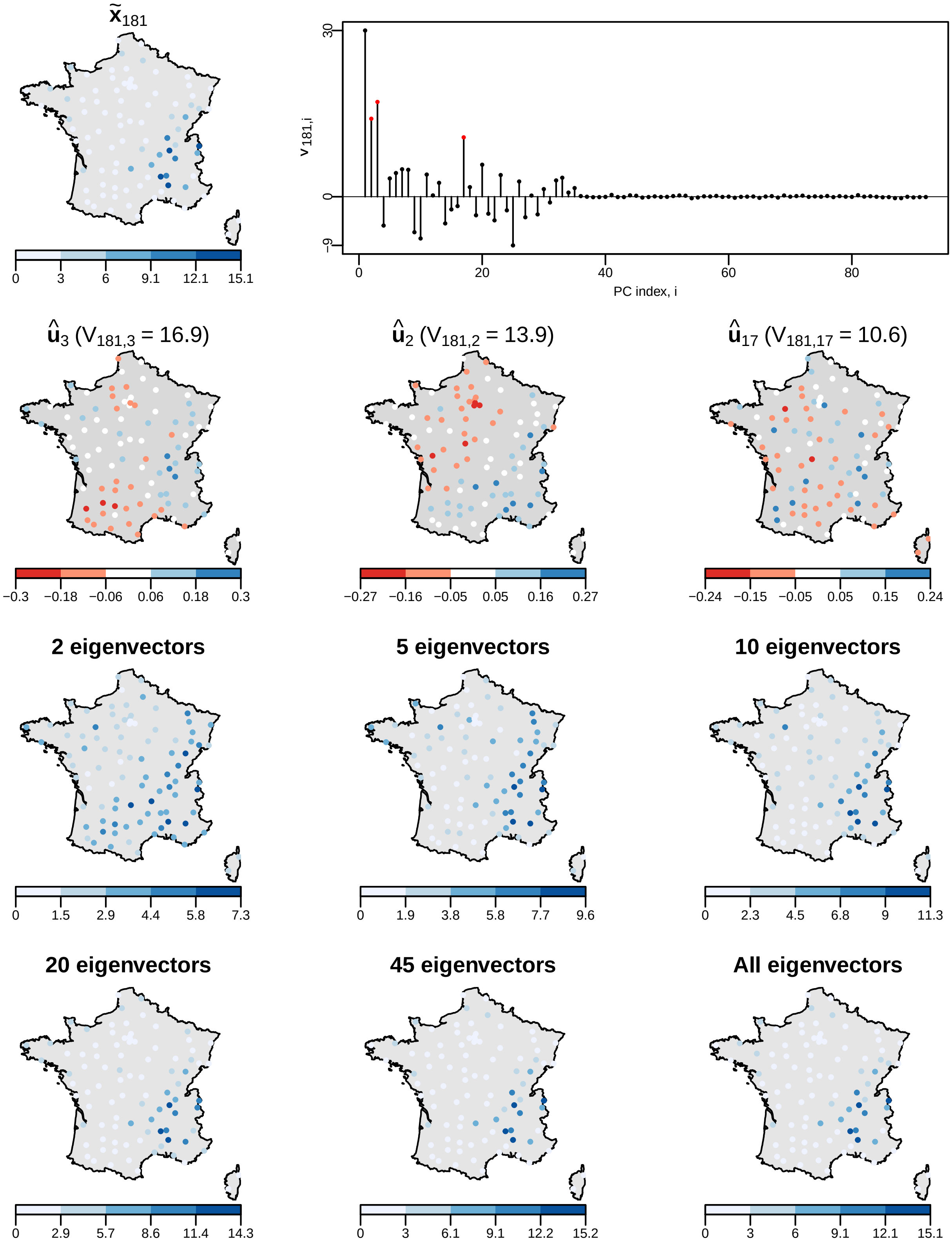

Figure 4.7: Exploration of the extreme event at $t = 181$. Top left: the observed event (after marginal transformation). Top right: the components of $\boldsymbol{v}_{181}$, with the three biggest components in absolute value (excluding the first principal component) highlighted in red. Second row: the eigenvectors corresponding to the highlighted principal components. Third and fourth rows: reconstructions based on a limited number of the leading eigenvectors.

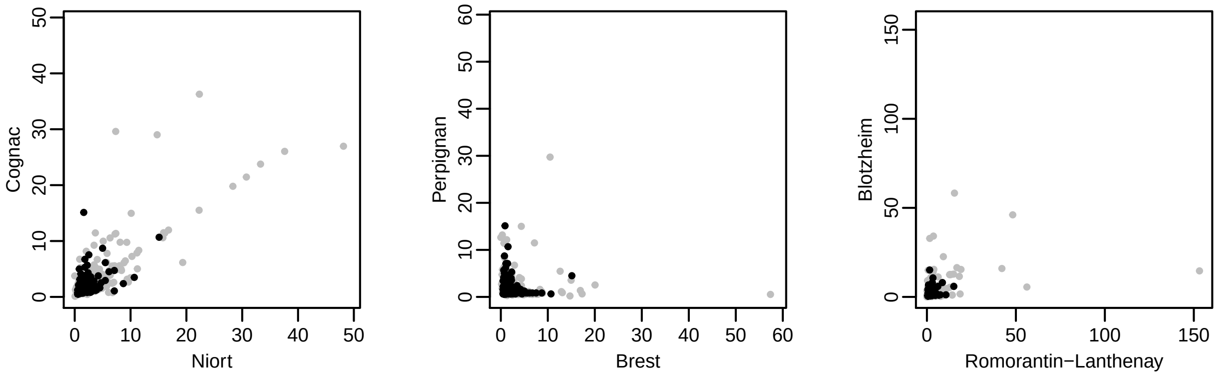

Figure 4.8: Pairwise plots for the simulated (grey) and observed (black) events at three pairs of stations with varying levels of extremal dependence. The simulated events are based on the 50-year hazard event set.

4.3 Generating hazard event sets

4.3.1 Fitting the mixture distribution for $\boldsymbol{Z}$

We assume that the first four eigenpairs capture the large-scale dynamics of heavy rainfall events. At this stage, this choice of $m = 4$ is somewhat arbitrary and the analysis in Section 4.2.3 indicates that m should actually be much larger - this will be investigated later in Section 4.3.3. From the principal components $\{\boldsymbol{v}_t : t = 1, \ldots, T\}$ we derive a set of observations $\{\boldsymbol{z_i} \in \mathbb{S}^5 : i = 1, \ldots, n_{\text{exc}}$ of $\boldsymbol{Z}$ by following the procedure described in Section 3.3.2, with r_V set as the empirical 85% quantile of $\{\|\boldsymbol{v}_t\|_2 : t = 1, \ldots, T\}$. The next step is to fit a Mises-Fisher mixture distribution for $\boldsymbol{Z}$. Following the suggestion of Rohrbeck and Cooley (2021), an estimate $\hat{\kappa} = 0.14$ for the kernel bandwidth is obtained using the `vmfkde.tune` function from the `Directional` package in R. The `rmixvmf` function from the same package is used to generate samples from this fitted distribution. Samples of $\widetilde{\boldsymbol{X}}$ can be derived from samples of $\boldsymbol{Z}$ as described earlier.

4.3.2 Analysing a single generated hazard event set

First, we simulate a single hazard event set corresponding to a 50-year period (i.e. $T = 600$ autumn weeks). Figure 4.8 compares the simulated and observed events at three pairs of stations. The dependencies for the simulated events reflect those of the observed events. This rough check suggests that the sampling algorithm is performing correctly. Figure 4.9 illustrates the three events with the largest size $\|\widetilde{\boldsymbol{x}}\|_2$. The left plot shows an event with extremely heavy rainfall in central and south-eastern France, including at Romorantin-Lanthenay (cf. the right-hand plot in Figure 4.8). The middle plot shows a very widespread event; the right plot shows an event where the rainfall intensity was especially high at a single site near Paris. This information is useful to practitioners for risk assessment/mitigation purposes.

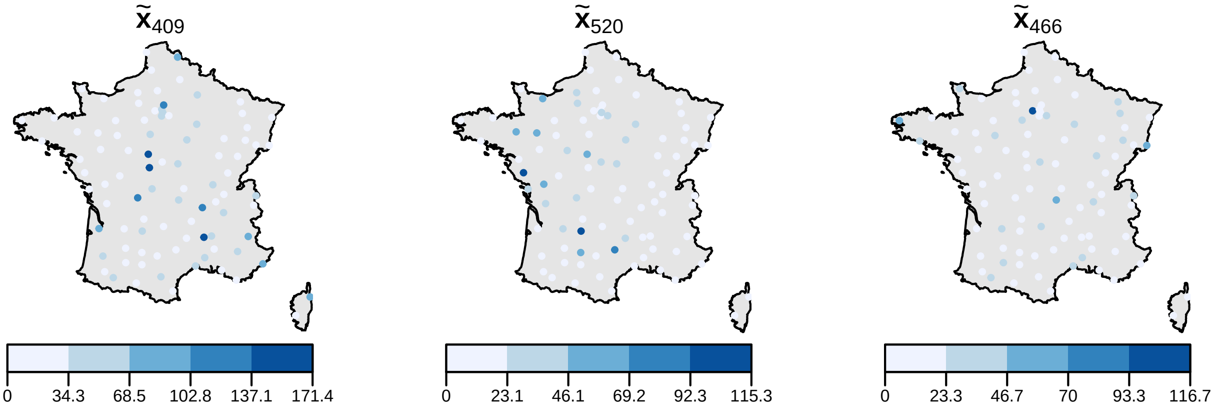
$\tilde{\mathbf{X}}_{409}$
0
34.3
68.5
102.8
137.1
171.4
$\tilde{\mathbf{X}}_{520}$
0
23.1
46.1
69.2
92.3
115.3
$\tilde{\mathbf{X}}_{466}$
0
23.3
46.7
70
93.3
116.7

Figure 4.9: Spatial representations of the three largest events from the simulated 50-year event set.

4.3.3 Choosing the hyperparameter m

The choice of m will impact the quality of the simulations. If m is too small, then the model will fail to capture the large-scale structure of extremes and the synthetic events will be unrealistic. If m is too large, then there is a risk of overfitting to the noise contained in the higher order eigenvectors, and the original statistical difficulty of estimating a high-dimensional distribution with a small effective sample size has not been avoided. This raises the question of how to select m, which has not been addressed thus far. We now explore two possible approaches for doing this.

The first method is inspired by the model fitting process in Fix et al. (2021). They estimate the parameter ρ of a spatial autoregressive (SAR) model by minimising the discrepancy $\|\Sigma(\rho) - \hat{\Sigma}\|_{\mathrm{F}}$ between $\Sigma(\rho)$, the theoretical TPDM under their model, and $\hat{\Sigma}$, the TPDM estimated from the original data. Here $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius matrix norm, given by

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_i \sum_j |a_{ij}|^2} \equiv \sqrt{\mathrm{trace}(A^\star A)}.$$

The process is as follows. Fix m and simulate a set of events. Then, compute the TPDM estimate $\hat{\Sigma}(m)$ using the simulated data and calculate the Frobenius distance $\|\hat{\Sigma}(m) - \hat{\Sigma}\|_{\mathrm{F}}$. This process is repeated for 1,000 hazard event sets and we take the median. The results (along with 95% bootstrap confidence intervals) are shown in Figure 4.10 (left plot). The graph exhibits a U-shape which agrees with our earlier hypothesis that m should be 'not too small' or 'not too large'. Here, $m \approx 12$ appears optimal.

The second approach follows an identical process, except now we estimate $\hat{\Sigma}(m)$ based on the set of m-eigenvector reconstructions of the observed events. From Figure 4.7 we know that the quality of the reconstruction improves as m increases so that we should expect $\|\hat{\Sigma}(m) - \hat{\Sigma}\|_{\mathrm{F}}$ to decrease with m, with diminishing returns as m gets large. The results are given in the right plot in Figure 4.10. We observe that the error is not strictly decreasing. This is because the reconstructions are optimal with respect to a metric that is different to $\|\hat{\Sigma}(m) - \hat{\Sigma}\|_{\mathrm{F}}$. The

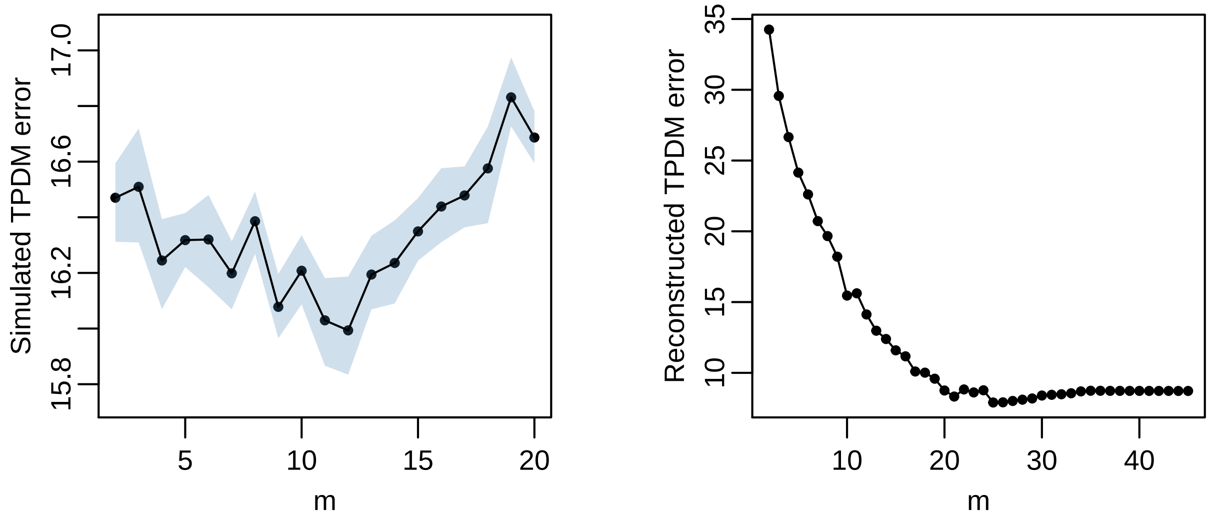

Figure 4.10: The error $\|\hat{\Sigma}(m) - \hat{\Sigma}\|_{\mathrm{F}}$ aginst m. Left: when $\hat{\Sigma}(m)$ is estimated simulated events; we take the median over 1,000 19-year event sets and compute 95% bootstrap confidence intervals. Right: when $\hat{\Sigma}(m)$ is estimated based on the set of m-eigenvector reconstructions.

plot suggests that taking $m \approx 20$ should be sufficient. This finding tallies with our earlier comments about the sequence of reconstructions in Figure 4.7. However, our two approaches for choosing m lead to different conclusions. This divergence, and other avenues to explore in future work, will be discussed further in the next chapter.

5
Future work

This concluding chapter will identify limitations in the existing methodology and outline some ideas for how these can be addressed in future work.

5.1 Choosing m in applications

The parameter m plays a crucial role in determining the realism and usefulness of the synthetic events generated by the sampling algorithm of Rohrbeck and Cooley (2021). However, a good criterion for selecting m has not been established thus far. The standard approach is to perform a spectral analysis and choose m such that $\sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{d} \lambda_i$ is acceptably close to 1. We seek more sophisticated approaches. The key question is: what is the best way to measure the agreement between two datasets of extreme events? Future work will aim to formulate a suitable notion of error upon which to base a selection criterion for m.

An important question to address is whether we should be considering reconstruction error or simulation error. The analysis in Section 4.3.3 suggests these notions of error do not coincide; this should be confirmed through theoretical work and further numerical studies. Within each of these categories, there are several notions of error. For example, we could consider quantities such as:

1. The event reconstruction error, $\sum_{t=1}^{T} \|\hat{\boldsymbol{x}}_t(m) - \widetilde{\boldsymbol{x}}_t\|$, where $\hat{\boldsymbol{x}}_t(m)$ is the m-eigenvector reconstruction of the observed event $\widetilde{\boldsymbol{x}}_t$;
2. The empirical conditional risk, see (3.4);
3. The TPDM error, $\|\hat{\Sigma}(m) - \hat{\Sigma}\|$, where $\hat{\Sigma}(m)$ is the TPDM estimated from a set of reconstructed/simulated events;
4. The χ error, $\|\hat{\chi}(m) - \hat{\chi}\|$, where $\hat{\chi}(m)$ is the matrix of tail correlation coefficients estimated from a set of reconstructed/simulated events;

The first measure is unlikely to be useful, because it will be distorted by the inaccurate magnitudes of the reconstructed events, whereas our focus is on capturing the dependence structure. The tail correlation measure χ is more robust than the TPDM in terms of its sensitivity to very extreme events in the dataset, so the fourth measure may be superior to the third. Some of these measures may be susceptible to rewarding overfitted models. A simulated event set that closely matches the observed event set will achieve a low error but is not useful in practice.

5.2 TPDM estimators

Recall from Section 4.2.2 that there are two different TPDM estimators being used in the literature. The vector-based estimator is a much more natural estimator, in the sense that it satisfies the theoretical properties of the TPDM, but it may perform poorly if extremal behaviour is highly localised. This trade-off is somewhat unsatisfactory and hints at a broader underlying issue. Therefore, we propose a critical comparison of the TPDM estimators and exploring ways they can be improved.

5.2.1 Bias correction

A first step for improving the TPDM estimator is to reduce bias. Threshold-based estimators are known to overestimate dependence between weakly dependent variables (Huser et al. 2016). Fix et al. (2021) address this by imposing that extremal dependence is close to zero at large distances. This assumption is reasonable in some applications (e.g. extreme precipitation on a large spatial domain) but not in others. For example, in an analysis of extreme river flow, two flow-connected sites may exhibit extremal dependence even if they are far apart, so it is more appropriate to correct bias based on the graph structure of the river network. However, this is counter to the spirit of our methods, which are supposed to be data-driven without the need to incorporate prior domain knowledge. Here we may turn to the wider extremes literature. The techniques described in Ledford (1996) might provide a useful starting point. They use so-called censored estimators to deal with the issues encountered when estimating the dependence between variables that are near independence. On the basis of results obtained via simulation studies, Huser et al. (2016) also recommend censored estimators, as well as choosing higher thresholds.

5.2.2 Permitting asymptotic independence

A similar, but separate, issue is to generalise extremal PCA so that it doesn't assume asymptotic dependence. Asymptotic independence is a degenerate case within the multivariate regular variation framework that underlies current methods, which cannot be attained from finite samples. We now outline some possible ways this could be achieved. Future work will involve testing and comparing these methods and/or devising some other new methodology.

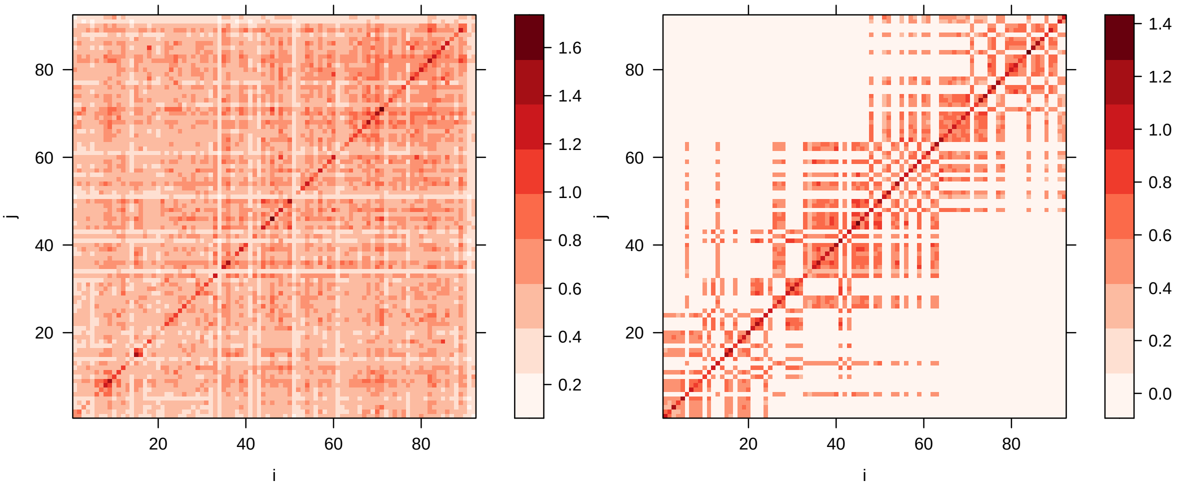

Figure 5.1: Left: the standard vector-based TPDM estimate for the French rainfall data. Right: the cluster vector-based TPDM estimate obtained after performing PAM clustering with F-madogram distances and $K = 5$.

An initial idea is to determine whether extremal PCA could be adapted to work with an alternative notion of regular variation called sparse regular variation (Meyer and Wintenberger 2021). Sparse regular variation redefines the angular component from $\boldsymbol{X}/\|\boldsymbol{X}\|$ to $\pi(\boldsymbol{X}/t)$, where t is the radial threshold and π is the Euclidean projection onto the unit simplex. Unlike with the self-normalised vector, the projected points need not lie in the interior of the simplex. Therefore, sparse regular variation better captures the sparse structure of H and permits asymptotic independence. Unfortunately, the independence between the radial and angular components is lost (though the dependence relation is known), so adapting extremal PCA for this framework may be complicated.

A second approach to consider is employ clustering before estimating the TPDM. Specifically, we partition the components of $\boldsymbol{X}$ into clusters, estimate the TPDM within each cluster separately, and then combine the cluster-based TPDMs to form the overall TPDM estimate. The inter-cluster entries of $\hat{\Sigma}$ are set equal to zero, meaning that components in different clusters are assumed to be asymptotically independent. Remark 3.3 in Fomichov and Ivanovs (2020), which alludes to links between spherical clustering and extremal PCA, may provide useful insight into the underlying theory of this idea. Figure 5.1 illustrates the result when this process is applied to the French rainfall data. The left plot shows the standard vector-based TPDM estimate we computed earlier, for reference. The right-hand TPDM is computed after performing clustering using the methodology of Bernard et al. (2013), i.e. the PAM algorithm with F-madogram distances and $K = 5$. Introducing this additional step means there are now two hyperparameters (K and m) to be tuned.

A third possible approach is very similar to the previous method, except the clustering step is replaced with another method for detecting the dependence structure. For example, there are a range of existing methods for detecting the faces of the unit sphere charged with

H-mass (Goix et al. 2017; Simpson et al. 2019; Meyer and Wintenberger 2021); the TPDM can then be estimated separately on each face.

5.3 Performing extremal PCA with large m

In their study of extreme rainfall in the UK, Rohrbeck and Cooley (2021) found that six extremal principal components were sufficient to describe large-scale extremal behaviour. In other applications, such as the French rainfall data, a larger number may be required. In such cases where drastic dimension reduction is not feasible, we may be need to perform extremal PCA with moderate or large m.

One approach is to explore whether there is any structure in the extremal principal components that can be exploited. We know that the extremal principal components are not asymptotically independent, but they do satisfy a property that implies balance between quadrants (Cooley and Thibaud 2019, Section 5). It is worth investigating whether this or some other structure can be exploited. As a starting point, we could try applying one of the clustering or face detection techniques to identify groups of extremal principal components with strong dependence, and then fit submodels to each of these groups.

Another approach is to explore alternatives to kernel density estimation (KDE) when fitting the flexible model for the leading components of H_V. KDE performance tends to worsen in high dimensions (Wang and Scott 2019), so we will consult the literature to find suitable alternative models for spherical data. The semi-parametric mixture models (which become fully-parametric when the number of mixture components is fixed) are potential candidates.

5.4 Uncertainty quantification

At present, the methodology does not include any considerations for uncertainty quantification. Uncertainty arises at several occasions in the modelling process, such as the estimation of the marginal distributions, the estimation of the TPDM, and fitting the KDE for $\boldsymbol{Z}$. Initially, we will focus on quantifying uncertainty in the TPDM and studying how it propagates through to the simulations. A natural way to estimate uncertainty is via a non-parametric bootstrap procedure. Suppose we are given a set of n observations of a random vector $\boldsymbol{X}$. For simplicity, we would initially consider a case where the marginal transformation has already been performed, and the true distribution of $\boldsymbol{X}$ is known, e.g. d-variate max-stable Hüsler-Reiss. We then sample with replacement from the original set of observations, estimate the TPDM, and generate a set of samples of $\boldsymbol{X}$. Then one can study the sampling distribution of the TPDM estimates and estimate the bias and variance. Moreover, if we have a way of estimating the optimal m (Section 5.1), we can also study variation in the estimates $\hat{m}$ - Fix et al. (2021) employ this approach to derive confidence intervals for the SAR parameter ρ. The procedure of computing a large number of bootstrapped TPDMs may lead to better approximations for the distribution of $\boldsymbol{X}$ (this can be checked in a simulation setting) - if this is the case, it opens up another route for advancing the methodology.

Appendices

A

Simulating d-variate max-stable Hüsler-Reiss data

A.1 Framework and assumptions

The unit sphere on the non-negative orthant S_+^{d-1} can be partitioned into $2^d - 1$ subspaces, called faces, given by

$$\mathcal{E}_I = \{\boldsymbol{\theta} \in \mathbb{S}_+^{d-1} : \theta_i > 0 \, \forall i \in I, \, \theta_j = 0 \, \forall j \notin I\}, \tag{A.1}$$

for $I \in \mathcal{P}(V) \setminus \emptyset$, where $\mathcal{P}(V)$ is the power set of $V = \{1, \ldots, d\}$. The face $\mathcal{E}_I$ being charged with H-mass, i.e. $H(\mathcal{E}_I) > 0$, indicates that the group of components $(X_i : i \in I)$ may be concomitantly extreme while the components $(X_i : i \notin I)$ are much smaller.

We consider the very simple scenario where there exists a partition of the index set such that asymptotic independence is present between groups but not necessarily within groups. Formally, assume there exists $2 \leq k \leq d$ and a partition $I_1 \cup \cdots \cup I_k = V$ such that the faces $\mathcal{E}_{I_1}, \ldots, \mathcal{E}_{I_k}$ form the support of the angular measure. Without loss of generality, assume that all indices in I_l are smaller than those in I_m for all $1 \leq l < m \leq k$. Then the faces $\mathcal{E}_{I_1}, \ldots, \mathcal{E}_{I_k}$ can be specified simply by providing their dimensions $d_1, \ldots, d_k$, where $d_i = |I_i|$ for $i = 1, \ldots, k$. Finally, define face labels $C_1, \ldots, C_d$ such that $C_i = m$ indicates that $i \in I_m$.

The random vector of interest, $\boldsymbol{X}$, is taken to have a d-variate max-stable Hüsler-Reiss (HR) distribution (Engelke, Malinowski et al. 2015). The dependence structure is parametrised by a conditionally negative definite matrix $\Gamma \in \mathbb{R}^{d \times d}$, called the variogram. The strength of dependence between X_i and X_j is controlled by Γ_{ij} via the relation $\chi_{ij} = 2\bar{\Phi}(\sqrt{\Gamma_{ij}}/2)$, where $\bar{\Phi}$ is the survival function of the standard normal distribution. The HR distribution captures all levels of dependence, from complete asymptotic dependence for $\Gamma_{ij} = 0$ to asymptotic independence for $\Gamma_{ij} \to \infty$.

A.2 Simulation methodology

Fomichov and Ivanovs (2020) describe a procedure for generating such data with $k = 2$ faces. We extend the algorithm for arbitrary $2 \leq k \leq d$ as follows:

1. Generate a collection of independent random vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_d \in \mathbb{R}^d$ whose components are independent, identically distributed Pareto random variables with shape parameter 2.5.
2. Fix $L \in \mathbb{R}$ a large constant and $\beta > 0$. Generate the variogram by setting

$$\Gamma_{ij} = \begin{cases} \frac{\beta}{d} \left\| \boldsymbol{h}_i - \boldsymbol{h}_j \right\|_2^2, & \text{if } C_i = C_j \\ L, & \text{if } C_i \neq C_j. \end{cases}$$

3. Generate samples of $\boldsymbol{X} \sim \text{HüslerReiss}(\Gamma)$, e.g. using the `rmstable` function from the `graphicalExtremes` package in R.

The parameter β scales the distribution of the extremal dependence coefficients, thereby controlling the strength of the within-group extremal dependence. Setting $\Gamma_{ij} = L$ enforces asymptotic independence between groups, since $\chi_{ij} = 2\bar{\Phi}(\sqrt{L}/2) \approx 0$.

Bibliography

Bador, Margot et al. (2017). 'Future Summer Mega-Heatwave and Record-Breaking Temperatures in a Warmer France Climate'. In: *Environmental Research Letters* 12.7, p. 074025. DOI: 10.1088/1748-9326/aa751c.

Beirlant, Jan et al. (2004). *Statistics of Extremes: Theory and Applications.* Wiley Series in Probability and Statistics. DOI: 10.1002/0470012382.

Bernard, Elsa et al. (2013). 'Clustering of Maxima: Spatial Dependencies among Heavy Rainfall in France'. In: *Journal of Climate* 26.20, pp. 7929–7937. DOI: 10.1175/JCLI-D-12-00836.1.

Boldi, M.-O. and A. C. Davison (2007). 'A Mixture Model for Multivariate Extremes'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2, pp. 217–229. DOI: 10.1111/j.1467-9868.2007.00585.x.

Bracken, C. et al. (2015). 'Spatial Variability of Seasonal Extreme Precipitation in the Western United States'. In: *Journal of Geophysical Research: Atmospheres* 120.10, pp. 4522–4533. DOI: 10.1002/2015JD023205.

Carreau, J., P. Naveau and L. Neppel (2017). 'Partitioning into Hazard Subregions for Regional Peaks-over-Threshold Modeling of Heavy Precipitation'. In: *Water Resources Research* 53.5, pp. 4407–4426. DOI: 10.1002/2017WR020758.

Chautru, Emilie (2015). 'Dimension Reduction in Multivariate Extreme Value Analysis'. In: *Electronic Journal of Statistics* 9.1, pp. 383–418. DOI: 10.1214/15-EJS1002.

Chiapino, Maël and Anne Sabourin (2017). 'Feature Clustering for Extreme Events Analysis, with Application to Extreme Stream-Flow Data'. In: *New Frontiers in Mining Complex Patterns.* Ed. by Annalisa Appice et al. Vol. 10312. Lecture Notes in Computer Science. Springer International Publishing, pp. 132–147. DOI: 10.1007/978-3-319-61461-8_9.

Chiapino, Maël, Anne Sabourin and Johan Segers (2018). *Identifying Groups of Variables with the Potential of Being Large Simultaneously.* arXiv: 1802.09977 [stat].

Coles, Stuart (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer Series in Statistics. Springer London. DOI: 10.1007/978-1-4471-3675-0.

Cooley, D and E Thibaud (2019). 'Decompositions of Dependence for High-Dimensional Extremes'. In: *Biometrika* 106.3, pp. 587–604. DOI: 10.1093/biomet/asz028.

de Carvalho, Miguel and Anthony C. Davison (2014). 'Spectral Density Ratio Models for Multivariate Extremes'. In: *Journal of the American Statistical Association* 109.506, pp. 764–776. DOI: 10.1080/01621459.2013.872651.

Dhillon, Inderjit S (2001). 'Concept Decompositions for Large Sparse Text Data Using Clustering'. In: *Machine Learning* 42, p. 33.

Drees, Holger and Anne Sabourin (2021). 'Principal Component Analysis for Multivariate Extremes'. In: *Electronic Journal of Statistics* 15.1, pp. 908–943. DOI: 10.1214/21-EJS1803.

Einmahl, John H. J., Fan Yang and Chen Zhou (2020). 'Testing the Multivariate Regular Variation Model'. In: *Journal of Business & Economic Statistics*, pp. 1–13. DOI: 10.1080/07350015.2020.1737533.

Engelke, Sebastian and Jevgenijs Ivanovs (2021). ‘Sparse Structures for Multivariate Extremes’. In: *Annual Review of Statistics and Its Application* 8.1, pp. 241–270. DOI: 10.1146/annurev-statistics-040620-041554.

Engelke, Sebastian, Alexander Malinowski et al. (2015). ‘Estimation of Hüsler-Reiss Distributions and Brown-Resnick Processes’. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.1, pp. 239–265. DOI: 10.1111/rssb.12074.

Fisher, R. A. and L. H. C. Tippett (1928). ‘Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample’. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2, pp. 180–190. DOI: 10.1017/S0305004100015681.

Fix, Miranda J., Daniel S. Cooley and Emeric Thibaud (2021). ‘Simultaneous Autoregressive Models for Spatial Extremes’. In: *Environmetrics* 32.2. DOI: 10.1002/env.2656.

Fomichov, V. and J. Ivanovs (2020). *Detection of Groups of Concomitant Extremes Using Clustering*. arXiv: 2010.12372 [math, stat]. URL: http://arxiv.org/abs/2010.12372.

Goix, Nicolas, Anne Sabourin and Stephan Clémençon (2017). ‘Sparse Representation of Multivariate Extremes with Applications to Anomaly Detection’. In: *Journal of Multivariate Analysis* 161, pp. 12–31. DOI: 10.1016/j.jmva.2017.06.010.

Gudendorf, Gordon and Johan Segers (2010). ‘Extreme-Value Copulas’. In: *Copula Theory and Its Applications*. Ed. by Piotr Jaworski et al. Vol. 198. Lecture Notes in Statistics. Springer Berlin Heidelberg, pp. 127–145. DOI: 10.1007/978-3-642-12465-5_6.

Gumbel, E J (1960). ‘Bivariate Exponential Distributions’. In: *Journal of the American Statistical Association* 55.292, pp. 698–707.

Hall, Peter, G.S. Watson and Javier Cabrera (1986). ‘Kernel Density Estimation with Spherical Data’. In: *Biometrika* 74.4, pp. 751–762. DOI: 10.2307/2336469.

Hanson, Timothy E., Miguel de Carvalho and Yuhui Chen (2017). ‘Bernstein Polynomial Angular Densities of Multivariate Extreme Value Distributions’. In: *Statistics & Probability Letters* 128, pp. 60–66. DOI: 10.1016/j.spl.2017.03.030.

Haug, Stephan, Claudia Klüppelberg and Gabriel Kuhn (2009). ‘Dimension Reduction Based on Extreme Dependence’. In: p. 20.

Huser, Raphaël, Anthony C. Davison and Marc G. Genton (2016). ‘Likelihood Estimators for Multivariate Extremes’. In: *Extremes* 19.1, pp. 79–103. DOI: 10.1007/s10687-015-0230-4.

Hüsler, Jürg and Rolf-Dieter Reiss (1989). ‘Maxima of Normal Random Vectors: Between Independence and Complete Dependence’. In: *Statistics & Probability Letters* 7.4, pp. 283–286. DOI: 10.1016/0167-7152(89)90106-5.

James, Gareth et al. (2021). *An Introduction to Statistical Learning*. 2nd ed. Springer.

Janßen, Anja and Phyllis Wan (2020). ‘K-Means Clustering of Extremes’. In: *Electronic Journal of Statistics* 14.1, pp. 1211–1233.

Jiang, Yujing, Daniel Cooley and Michael F. Wehner (2020). ‘Principal Component Analysis for Extremes and Application to U.S. Precipitation’. In: *Journal of Climate* 33.15, pp. 6441–6451. DOI: 10.1175/JCLI-D-19-0413.1.

Kaufman, Leonard and Peter J. Rousseeuw (1990). *Finding Groups in Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc. DOI: 10.1002/9780470316801.

Larsson, Martin and Sidney I. Resnick (2012). ‘Extremal Dependence Measure and Extremogram: The Regularly Varying Case’. In: *Extremes* 15.2, pp. 231–256. DOI: 10.1007/s10687-011-0135-9.

Ledford, A. (1996). ‘Statistics for near Independence in Multivariate Extreme Values’. In: *Biometrika* 83.1, pp. 169–187. DOI: 10.1093/biomet/83.1.169.

Meyer, Nicolas and Olivier Wintenberger (2021). 'Sparse Regular Variation'. In: arXiv: 1907.00686.

Mhatre, Nehali and Daniel Cooley (2021). *Transformed-Linear Models for Time Series Extremes*. arXiv: 2012.06705 [stat].

Mornet, Alexandre et al. (2017). 'Wind Storm Risk Management: Sensitivity of Return Period Calculations and Spread on the Territory'. In: *Stochastic Environmental Research and Risk Assessment* 31.8, pp. 1977–1995. DOI: 10.1007/s00477-016-1367-7.

Reich, Brian J. and Benjamin A. Shaby (2019). 'A Spatial Markov Model for Climate Extremes'. In: *Journal of Computational and Graphical Statistics* 28.1, pp. 117–126. DOI: 10.1080/10618600.2018.1482764.

Resnick, Sidney I. (1987). *Extreme Values, Regular Variation and Point Processes*. Red. by Thomas V. Mikosch, Sidney I. Resnick and Stephen M. Robinson. Springer Series in Operations Research and Financial Engineering. Springer New York. DOI: 10.1007/978-0-387-75953-1.

— (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering. Springer. 404 pp.

Rohrbeck, Christian and Daniel Cooley (2021). *Simulating Flood Event Sets Using Extremal Principal Components*. arXiv: 2106.00630 [stat]. URL: http://arxiv.org/abs/2106.00630.

Rohrbeck, Christian and Jonathan A. Tawn (2020). 'Bayesian Spatial Clustering of Extremal Behavior for Hydrological Variables'. In: *Journal of Computational and Graphical Statistics*, pp. 1–15. DOI: 10.1080/10618600.2020.1777139.

Rousseeuw, Peter J. (1987). 'Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis'. In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

Saunders, K. R., A. G. Stephenson and D. J. Karoly (2020). 'A Regionalisation Approach for Rainfall Based on Extremal Dependence'. In: *Extremes*. DOI: 10.1007/s10687-020-00395-y.

Schervish, Mark J. (1995). *Theory of Statistics*. Springer Series in Statistics. Springer New York. DOI: 10.1007/978-1-4612-4250-5.

Simpson, Emma S., Jennifer L. Wadsworth and Jonathan A. Tawn (2019). *Determining the Dependence Structure of Multivariate Extremes*. arXiv: 1809.01606 [stat].

Tawn, Jonathan A (1988). 'Bivariate Extreme Value Theory: Models and Estimation'. In: *Biometrika* 75.3, pp. 397–415.

Vignotto, Edoardo, Sebastian Engelke and Jakob Zscheischler (2021). 'Clustering Bivariate Dependencies of Compound Precipitation and Wind Extremes over Great Britain and Ireland'. In: *Weather and Climate Extremes* 32, p. 100318. DOI: 10.1016/j.wace.2021.100318.

Wackernagel, Hans (1995). *Multivariate Geostatistics: An Introduction with Applications*. Springer Berlin Heidelberg.

Wadsworth, J. L. (2016). 'Exploiting Structure of Maximum Likelihood Estimators for Extreme Value Threshold Selection'. In: *Technometrics* 58.1, pp. 116–126. DOI: 10.1080/00401706.2014.998345.

Wadsworth, J. L. and J. A. Tawn (2012). 'Likelihood-Based Procedures for Threshold Diagnostics and Uncertainty in Extreme Value Modelling: Extreme Value Modelling'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3, pp. 543–567. DOI: 10.1111/j.1467-9868.2011.01017.x.

Wang, Zhipeng and David W. Scott (2019). 'Nonparametric Density Estimation for High-Dimensional Data - Algorithms and Applications'. In: *WIREs Computational Statistics* 11.4. DOI: 10.1002/wics.1461. arXiv: 1904.00176.