Extensions and Applications of the Tail Pairwise Dependence Matrix

Matthew Pawley

October 19, 2024

Table of contents

List of Figures

List of Tables

Preface

Draft thesis of Matthew Pawley, created on October 19, 2024.

1 Introduction

1.1 Motivation

1.2 Thesis aims and outline

- Summarise general idea of the thesis.
- Chapter 2: introduction to key concepts of EVT; define TPDM, describe its properties, and review its applications so far; explain and demonstrate bias issue when dependence is weak.
- Chapter 3: EVA Data Challenge
- Chapter 4: changing dependence
- Chapter 5: compositional perspectives
- Chapter 6: shrinkage TPDM, sparse/robust methods etc. to handle the bias issue
- Chapter 7: summary, discussion and outlook

2 Literature review

2.1 Univariate extreme value theory

2.1.1 Block maxima and the generalised extreme value (GEV) distribution

Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed, continuous random variables with distribution function F. For $n \geq 1$, define the random variable

$$M_n := \max(X_1, \ldots, X_n) = \bigvee_{i=1}^{n} X_i. \tag{2.1}$$

The exact distribution of M_n is given by

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, \ldots X_n \leq x) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x) = F^n(x), \qquad (x \in \mathbb{R}).$$

This result is not particularly useful in practice, where F is typically unknown. Instead, we study the limiting behaviour of F^n as $n \to \infty$. Clearly the asymptotic distribution of M_n is degenerate, since $M_n \xrightarrow{p} x_F := \sup\{x : F(x) < 1\}$, the (possibly infinite) upper end-point of F. However, the Extremal Types Theorem states that, after suitable rescaling, there are three classes of non-degenerate asymptotic distribution (CITE).

Theorem 2.1. *Suppose there exist real sequences* $\{a_n > 0\}$ *and* $\{b_n \in \mathbb{R}\}$ *and a non-degenerate distribution function* G *such that*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{d} G(x), \qquad (n \to \infty). \tag{2.2}$$

Then G *belongs to one of three parametric families: Gumbel, Fréchet or negative Weibull.*

When (2.2) holds, we say that F lies in the maximum domain of attraction (MDA) of G. The three families are unified by the Generalised Extreme Value (GEV) distribution. Its distribution function is

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, \tag{2.3}$$

where $[x]_+ := \max(0, x)$ denote the positive part of x. The parameters $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$ are called the location, scale, and shape, respectively. The sign of the shape parameter determines the sub-class that G belongs to: $\xi > 0$ corresponds to the heavy-tailed Fréchet class, $\xi = 0$ (with (2.2) interpreted as $\xi \to 0$) corresponds the exponential-tailed Gumbel class, and $\xi < 0$ the negative Weibull class, which has a finite upper limit.

The GEV distribution is used to model the upper tail of X via the block maxima approach (CITE). Let $x_1, \ldots, x_n$ denote independent observations of $X_1, \ldots, X_n$. The data are partitioned into finite blocks of size m. Provided m is sufficiently large, the maximum observation in each block is approximately GEV distributed by Theorem 2.1. Once the block-wise maxima have been extracted, estimates of the GEV parameters may be obtained, e.g. by maximum likelihood inference. The performance of the fitted model is sensitive to the choice of block size. Selection of the tuning parameter m requires managing a bias-variance trade-off. If the blocks are too small, then the underlying asymptotic approximation may not be valid and the maxima may not be representative as extreme events, biasing the estimates. Taking larger blocks reduces the amount of data available for inference, resulting in noisier estimation of the GEV parameter estimates.

2.1.2 Threshold exceedances and the generalised Pareto distribution (GPD)

The block maxima procedure is considered inefficient, because it fails to exploit all the available information. Each block is summarised by a (single) maximum value, even if it contains other 'extreme' events that might be informative for the tail. The intimately related peaks-over-threshold method makes better use of the available data. If X is in the maximum domain of attraction of a GEV(μ, σ, ξ) distribution, then

$$\lim_{u\to\infty} \mathbb{P}(X - u > x \mid X > u) = \left[1 + \frac{\xi x}{\tilde{\sigma}}\right]_+^{-1/\xi}, \qquad (x > 0), \tag{2.4}$$

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ (CITE). The limiting conditional distribution is called the generalised Pareto distribution (GPD). The GPD describes the distribution of excesses over a high threshold. Given observations $x_1, \ldots, x_n$, the peaks-over-threshold method assumes that exceedances of some pre-specified high threshold $u > 0$ are approximately GPD distributed. Maximum likelihood or Bayesian inference procedures may be used to estimate the GPD parameters $\bar{\sigma}, \xi$. Threshold selection is subject to similar considerations as for the block size. Picking a low threshold risks model misspecification, causing bias in the fitted model. Choosing a high threshold directly reduces the number of threshold exceedances, increasing the uncertainty in the parameter estimates. Various diagnostics and procedures have been proposed to aid with this choice. Many approaches rely on inspecting diagnostic plots, such as mean residual life (MRL) plots (CITE) and parameter stability plots (CITE). Automated selection procedures aim to remove subjectivity by optimising with respect to some criterion. These include change-point methods (CITE Wadsworth 2016), cross-validation in a Bayesian framework (CITE Northrop et al. 2017), and minimising expected quantile discrepancies (CITE Murphy and Tawn 2024).

2.1.3 Non-stationary extremes

The block-maxima and peaks-over-threshold methods as presented above assume that the data are stationary over the observation period. In environmental applications, climate change threatens the validity of this assumption, with changes in the frequency and intensity of extreme weather events (CITE). Non-stationary models accommodate temporal dependence by allowing parameters to vary over time or in relation to covariates. For example, CITE Vanem 2015 incorporate trends into the GEV location and scale parameters by specifying

$$\mu(t) = \mu_0 + \mu_1 t, \qquad \sigma(t) = \exp(\sigma_0 + \sigma_1 t).$$

If the parameters μ_1 and σ_1 are significantly different from zero, it suggests the data exhibit non-stationarity. In principle the shape parameter may be extended analogously. Often the shape parameter is assumed constant because is notoriously difficult to estimate accurately and results (quantiles, return periods, etc.) are very sensitive to changes in its sign. *CITE further papers or a review?*

2.2 Multivariate extreme value theory

Multivariate extreme value theory (MEVT) generalises the study of extreme events from univariate to multivariate settings. Understanding the joint tail behaviour of several variables is critical in various fields. In environmental science, practitioners are tasked with assessing the risk of compound extreme events involving several variables. For example, the impact of drought – defined by the IPCC (CITE) as a prolonged period of low precipitation – is exacerbated by high temperatures. Similarly, extreme rainfall occurring simultaneously across multiple locations may lead to a widespread flood event. In finance, investors seek to diversify their portfolio to mitigate against the risk of simultaneous extreme losses across multiple assets. Each of these examples calls for a statistical analysis of the joint tail distribution of some random vector.

2.2.1 Componentwise maxima

Consider a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with unknown joint distribution function F, meaning

$$F(\boldsymbol{x}) := \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d),$$

for any $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be a sequence of independent copies of $\boldsymbol{X}$. The notion of 'extremes' or a 'maximum' becomes subjective in the multivariate setting, because $\mathbb{R}^d$ is not an ordered set. One possibility is to define the maximum component-wise as

$$\boldsymbol{M}_n := \left(\bigvee_{i=1}^{n} X_{i1}, \ldots, \bigvee_{i=1}^{n} X_{id} \right).$$

We say that F lies in the multivariate MDA of a non-degenerate distribution G if there exist $\mathbb{R}^d$-valued sequences $\{\boldsymbol{a}_n > \boldsymbol{0}\}$ and $\{\boldsymbol{b}_n \in \mathbb{R}^d\}$ such that

$$\mathbb{P}\left(\frac{\boldsymbol{M}_n - \boldsymbol{b}_n}{\boldsymbol{a}_n} \leq \boldsymbol{x} \right) \xrightarrow{d} G(\boldsymbol{x}), \qquad (n \to \infty). \tag{2.5}$$

Applying Theorem 2.1 to the marginal components reveals that the margins of G follow a univariate GEV distribution. The crucial difference to the univariate setting is that now the limit (joint) distribution G does *not* admit a parametric representation. The

inherently challenging nature of MEVT largely stem from this fact. The problem of estimating/modelling G is usually split into two (sequential) steps. First, one models the margins to describe the extreme behaviour of each variable individually (using univariate EVT). Then, one standardises to common margins and models the extremal dependence structure, i.e. the inter-relationships between extremes across multiple variables. Copula theory provides a rigorous justification for this two-step process.

2.2.2 Copulae and marginal standardisation

In multivariate statistics, Sklar's theorem allows for the separation of the marginal distributions of variables from their joint dependence structure through the use of a copula. It states that any multivariate distribution can be expressed as a combination of individual marginal distributions and a copula that captures the dependence between them.

Theorem 2.2. *Suppose $\boldsymbol{X} = (X_1, \ldots, X_d)$ has joint distribution function F and continuous marginal distributions $X_i \sim F_i$ for $i = 1, \ldots, d$. Then there exists a unique copula C such that*

$$F(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_d(x_d)\right). \tag{2.6}$$

The copula C characterises the dependence structure of the variables, and represents the distribution function of $\boldsymbol{X}$ after transforming to standard uniform margins. Uniform margins are a standard choice in multivariate statistics, but copulae may be defined with alternative marginal distributions. In extreme value theory, it is common to use Fréchet, exponential or Gumbel margins. The different choices accentuate particular features of the extreme values. For example, heavy-tailed Fréchet margins serve to highlight the most extreme values, while Gumbel or exponential margins are often favoured for conditional extremes modelling (CITE Heffernan and Tawn). Although the marginal distribution is an important modelling choice, ultimately all choices are valid/equivalent in the sense that monotonic transformations of the univariate marginals do not change the nature of tail dependence (Resnick 2007).

There are broadly two ways of performing the preliminary marginal standardisation. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d)$ has marginal distributions $X_i \sim F_i$ for $i = 1, \ldots, d$. If the functions

F_i are known, then the marginal distributions can be transformed to some common target distribution $F_\star$ via the probability integral transform:

$$X_i \mapsto F_\star^{-1}(F_i(X_i)) \sim F_\star, \qquad (i = 1, \ldots, d). \tag{2.7}$$

If the marginal distributions are unknown, as is usually the case, then F_i is replaced with some estimate $\hat{F}_i$ in (2.7). A standard choice for $\hat{F}_i$ is the empirical CDF (non-parametric), perhaps with GPD tails above a high threshold (semi-parametric). Examples of these two approaches can be found in Russell and Hogan (2018) and Rohrbeck and Cooley (2023), respectively. Throughout this thesis, uncertainty arising from estimation of the marginal distributions shall be neglected. Relaxing this assumption, as in Clémençon et al. (2023), represents an avenue for future work.

2.2.3 The exponent measure and angular measure

Suppose $\boldsymbol{X}$ is on unit Fréchet margins, that is

$$\mathbb{P}(X_i < x) = \exp(-1/x), \qquad (x > 0), \tag{2.8}$$

for $i = 1, \ldots, d$. This corresponds to a GEV distribution with $\mu = \sigma = \xi = 1$. The joint distribution G in (2.5) may be rewritten in the form

$$G(\boldsymbol{x}) = \exp(-V(\boldsymbol{x})), \tag{2.9}$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $x_i > 0$ for $i = 1, \ldots, d$. The exponent measure V is a function of the form

$$V(\boldsymbol{x}) = d \int_{\mathbb{S}_{+(1)}^{d-1}} \bigvee_{i=1}^{d} \left(\frac{\theta_i}{x_i}\right) \mathrm{d}H(\boldsymbol{\theta}). \tag{2.10}$$

Here

$$\mathbb{S}_{+(p)}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\|_p = 1\} \tag{2.11}$$

denotes the L_p-simplex in the non-negative orthant of $\mathbb{R}^d$ and the angular measure H is a probability measure on $\mathbb{S}^{d-1}_{+(1)}$ satisfying the moment constraints

$$\int_{\mathbb{S}^{d-1}_{+(1)}} \theta_i \, \mathrm{d}H(\boldsymbol{\theta}) = 1/d, \qquad (i = 1, \ldots, d). \tag{2.12}$$

Our notation for the simplex is borrowed from Fix et al. (2021). The exponent $d-1$ highlights the fact that the simplex is a $(d-1)$-dimensional set embedded in the d-dimensional space $\mathbb{R}^d$. The $+$ and (p) in the subscript convey that the set is restricted to the non-negative orthant and is with respect to the L_p-norm, respectively. The constraints on H arise due to tail equivalence of the margins. Functions G satisfying (2.9) are called multivariate extreme value distributions. If V is differentiable, then the density h of H exists in the interior and on the low-dimensional boundaries of the simplex. The relation between V and h is given by

$$h\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}\right) = -\frac{\|\boldsymbol{x}\|_1^{d+1}}{d} \frac{\partial^d}{\partial x_1 \cdots \partial x_d} V(\boldsymbol{x}). \tag{2.13}$$

The benefit of introducing the exponent and angular measures is that models for G may be specified in terms of V or H. The extremal dependence structure of $\boldsymbol{X}$ is completely characterised by H: the angular measure determines V via (2.10) and subsequently G via (2.9). Modelling the angular measure now becomes our primary focus.

2.2.4 Parametric multivariate extreme value models

The class of valid dependence structures is in direct correspondence to the infinite-dimensional class of valid measures H. This greatly hinders efforts to perform statistical inference: efficient estimation via likelihood inference, hypothesis testing, and inclusion of covariates immediately become unavailable. We may return to the parametric paradigm by postulating a suitable parametric sub-family. Ideally the chosen sub-family generates a wide class of valid dependence structures. A detailed review of popular models can be found in Gudendorf and Segers (2010).

There are several drawbacks to the parametric approach. Working with a parametric model instead of the general class runs the risk of model misspecification. Generating valid models is a challenging endeavour due to the moment constraints, resulting in models that are either overly simplistic or have unwieldy distribution functions and parameter

constraints. Striking a balance between flexibility and parsimony becomes especially in high dimensions (i.e. when d is large). For these reasons, parametric models are not a primary focus of this thesis. Nevertheless, we now review a small selection of models. These primarily feature as data-generating processes for our numerical experiments. Functionality for generating independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\boldsymbol{X}$ or $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim H$ based on the sampling algorithms formulated in Dombry et al. (2016) is provided in the R package `mev`.

2.2.4.1 Logistic-type models

One of the oldest and simplest multivariate extreme value models is the symmetric logistic distribution (Gumbel 1960).

Definition 2.1. The exponent measure of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ following the symmetric logistic distribution is

$$V(\boldsymbol{x}) = \left(\sum_{i=1}^{d} x_i^{-1/\gamma} \right)^{\gamma}, \qquad \gamma \in (0, 1]. \tag{2.14}$$

The single dependence parameter $\gamma \in (0, 1]$ characterises the strength of the association between all variables. Independence occurs when $\gamma = 1$ and the variables approach complete dependence as $\gamma \to 0$. All variables are exchangeable, since the distribution function is invariant under coordinate permutation. A flexible extension is the asymmetric logistic model of Jonathan A Tawn (1990). Greater control over the dependence structure is achieved by increasing the number of parameters.

Definition 2.2. The exponent measure of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ following the asymmetric logistic distribution is of the form

$$V(\boldsymbol{x}) = \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset} \left[\sum_{i \in \beta} \left(\frac{\theta_{i,\beta}}{x_i} \right)^{1/\gamma_\beta} \right]^{\gamma_\beta}, \qquad \begin{cases} \gamma_\beta \in (0, 1], & \\ \theta_{i,\beta} \in [0, 1], & \text{if } i \in \beta, \\ \theta_{i,\beta} = 0, & \text{if } i \notin \beta, \\ \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset} \theta_{i,\beta} = 1, & \end{cases} \tag{2.15}$$

where $\mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset$ denotes the set of non-empty subsets of $\{1,\ldots,d\}$.

The set of parameters $\{\gamma_\beta : \beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset\}$ control the dependence strength among the corresponding variables $\{X_i : i \in \beta\}$ in a similar way to the symmetric logistic model. The model's complexity arises from the set of asymmetry parameters $\boldsymbol{\theta}_\beta = (\theta_{i,\beta} : i \in \beta)$, which dictate the direction/composition of extreme events involving the variables $\{X_i : i \in \beta\}$. Further models can be generated by 'inverting' the logistic and asymmetric models. **The purpose of inverting is...**. When applied to the models described above, inversion yields the negative symmetric logistic model (Galambos 1975) and the negative asymmetric logistic model (Joe 1990), respectively.

Definition 2.3. The exponent measure of a random vector $\boldsymbol{X} = (X_1,\ldots,X_d)$ following the negative symmetric logistic distribution is

$$V(\boldsymbol{x}) = \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\})\setminus\emptyset} (-1)^{|\beta|+1} \left(\sum_{i\in\beta} x_i^\gamma\right)^{-1/\gamma}, \qquad \gamma > 0. \tag{2.16}$$

Definition 2.4. The exponent measure of a random vector $\boldsymbol{X} = (X_1,\ldots,X_d)$ following the negative asymmetric logistic distribution is

$$V(\boldsymbol{x}) = \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\})\setminus\emptyset} (-1)^{|\beta|+1} \left(\sum_{i\in\beta} x_i^\gamma\right)^{-1/\gamma}, \qquad \gamma > 0. \tag{2.17}$$

Other logistic-type models include the bilogistic Smith et al. (1990)] and negative bilogistic (Coles and J A Tawn 1994).

2.2.4.2 The Brown-Resnick process and Hüsler-Reiss distribution

The Brown-Resnick process of Brown and Resnick (1977) is a class of stochastic processes commonly used to model the extremal dependence structure of spatial phenomena, including rainfall (Davison et al. 2012), snow depths (Schellander and Hell 2018) and wind gusts (Oesting et al. 2017). It is naturally defined through a transformation of a Gaussian process

– a formal construction can be found in CITE Kabluchko et al. (2009). Let $\Omega \in \mathbb{R}^2$ be a spatial domain. Consider a Brown-Resnick process $\{X(\boldsymbol{s}) : \boldsymbol{s} \in \Omega\}$ with semi-variogram

$$\gamma(\boldsymbol{s}, \boldsymbol{s}') = (\|\boldsymbol{s} - \boldsymbol{s}'\|_2/\rho)^\kappa, \qquad \rho > 0, \kappa \in (0, 2]. \tag{2.18}$$

Semi-variograms of this form are called fractal semi-variograms and the associated process $\{X(\boldsymbol{s})\boldsymbol{s} \in \Omega\}$ is stationary and isotropic (Engelke, Malinowski, et al. 2015). Stationarity and isotropy mean that the statistical properties of the spatial process are invariant under translation and rotation. Specifically, the dependence between two sites only depends on the distance between them, not the direction or their position within the spatial domain. The parameters ρ and κ in (2.18) control the range and smoothness, respectively. The range parameter determines how quickly the dependence strength decreases over distance. The smoothness parameter governs the regularity of the process and affects its local behaviour.

Let $\boldsymbol{s}_i, \boldsymbol{s}_j \in \Omega$ be a pair of spatial locations and define random variables $X_i = X(\boldsymbol{s}_i)$ and $X_j = X(\boldsymbol{s}_j)$. The exponent measure of the bivariate random vectors (X_i, X_j) is (R. Huser and Davison 2013)

$$V(x_i, x_j) = \frac{1}{x_i}\Phi\left(\frac{a_{ij}}{2} + \frac{1}{a_{ij}}\log\frac{x_j}{x_i}\right) + \frac{1}{x_j}\Phi\left(\frac{a_{ij}}{2} + \frac{1}{a_{ij}}\log\frac{x_i}{x_j}\right), \tag{2.19}$$

where $a_{ij} = \sqrt{\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)}$. The stationary/isotropic nature of the underlying process is apparent because V depends on $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ only through $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2$.

Other things I could mention: Davison et al. (2012) apply BR to rainfall data, finding $1/2 < \kappa < 1$. *Although the Brown–Resnick processes are max-stable, the processes observed at a finite number of locations are also multivariate regularly varying.*

The Brown-Resnick process is intimately related to the Hüsler-Reiss distribution of Hüsler and Reiss (1989). The Hüsler-Reiss distribution is of fundamental importance in multivariate extremes: it has been labelled the Gaussian distribution for extremes (Engelke and Hitz 2019). In $d \geq 2$ dimensions the distribution is parametrised by a matrix $\Lambda = (\lambda_{ij}^2)_{1 \leq i,j \leq d}$

belonging to the class of symmetric, strictly conditionally negative definite matrices

$$\mathcal{D} := \left\{ M \in \mathbb{R}_+^{d\times d} \,:\, M = M^T,\, \operatorname{diag}(M) = \mathbf{0},\, \boldsymbol{x}^T M \boldsymbol{x} < 0 \,\forall \boldsymbol{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\} \text{ such that } \sum_{j=1}^{d} x_j = 0 \right\}.$$

The class of Hüsler-Reiss distributions is closed in the sense that if $\boldsymbol{X} = (X_1, \ldots, X_d)$ follows a Hüsler-Reiss distribution with parameter matrix Λ, then any random sub-vector (X_i, X_j) is also Hüsler-Reiss distributed with parameter λ_{ij}^2. This permits very flexible control over the pairwise dependence structure. The dependence between any pair of variables X_i and X_j can be adjusted by modifying the corresponding parameter λ_{ij}, subject to the constraint $\Lambda \in \mathcal{D}$. The finite-dimensional distribution of a Brown-Resnick process at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d$ is precisely the Hüsler-Reiss distribution with $\Lambda = (\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)/4)_{1\le i,j\le d}$ (Engelke, Malinowski, et al. 2015). Due to this link, the Hüsler-Reiss distribution may be parametrised in terms of its variogram matrix $\Gamma := 4\Lambda \in \mathcal{D}$ (Engelke and Jevgenijs Ivanovs 2021; Fomichov and Ivanovs 2023) and the exponent measure of (X_i, X_j) is given by (2.19) with a_{ij} replaced by $2\lambda_{ij}$.

2.2.4.3 The max-linear model

The final parametric model we consider is the max-linear (factor) model (Einmahl, Krajina, et al. 2012; Fougères et al. 2013; Yuen and Stoev 2014a). *Its exact origin is unclear, but it seems to stem from around these papers.* Max-linear models are a simple but flexible class possessing important theoretical properties. Any discrete angular measure concentrating on finitely many points corresponds to a max-linear model (Yuen and Stoev 2014a). Due to its flexibility and theoretical properties, the max-linear model has enjoyed widespread use across several areas of extremes, including clustering (Janßen and Wan 2020; Medina et al. 2021), graphical modelling for causal inference (Gissibl, Klüppelberg, and Lauritzen 2019; Gissibl and Klüppelberg 2018; Tran et al. 2021) and tail event probability estimation (Kiriliouk and Zhou 2022). In future sections/chapters, the max-linear model will be applied in more general settings where the marginal distributions are Fréchet with shape parameter $\alpha \ge 1$ and the angular measure is defined with respect to the L_α-norm on $\mathbb{R}^d$. In anticipation of this, the max-linear model is introduced in this more general setting. To revert to the setting established in the previous sections, the reader may simply take

$\alpha = 1$.

Definition 2.5. Let $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q) \in \mathbb{R}_+^{d\times q}$ for some $q \geq 1$. Assume that $\boldsymbol{a}_j \neq \boldsymbol{0}$ for all $j = 1, \ldots, q$ and each row has unit L_α-norm, i.e. $\sum_{j=1}^q a_{ij}^\alpha = 1$ for $i = 1, \ldots, d$. A random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with discrete probability angular measure

$$H(\cdot) = \frac{1}{\sum_{j=1}^q \|\boldsymbol{a}_j\|_\alpha^\alpha} \sum_{j=1}^q \|\boldsymbol{a}_j\|_\alpha^\alpha \delta_{\boldsymbol{a}_j/\|\boldsymbol{a}_j\|_\alpha}(\cdot) \tag{2.20}$$

is said to follow the max-linear model with parameter matrix A.

The row-wise unit-norm constraint on A results ensures the marginal components are Fréchet distributed with unit scale and shape α. Setting $\alpha = 1$, we see that (2.20) is a valid angular measure: for any $i = 1, \ldots, d$,

$$\int_{\mathbb{S}_{+(1)}^{d-1}} \theta_i \,\mathrm{d}H(\boldsymbol{\theta}) = \frac{1}{\sum_{j=1}^q \|\boldsymbol{a}_j\|_1} \sum_{j=1}^q \int_{\mathbb{S}_{+(1)}^{d-1}} \theta_i \|\boldsymbol{a}_j\|_1 \delta_{\boldsymbol{a}_j/\|\boldsymbol{a}_j\|_1}(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} = \frac{\sum_{j=1}^q a_{ij}}{\sum_{i=1}^d \sum_{j=1}^q a_{ij}} = \frac{1}{d}.$$

The number of free parameters is $d\times(q-1)$ and the order of the columns of A is inconsequential. The factors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q$ correspond to the possible directions that extremal observations may take. The column norms $\|\boldsymbol{a}_1\|_\alpha, \ldots, \|\boldsymbol{a}_q\|_\alpha$ determine the respective weights assigned to these directions. There is a direct correspondence between the class of discrete angular measure placing mass on $q < \infty$ points and the class of max-linear random vectors with q factors (Yuen and Stoev 2014a). Moreover, the class of angular measures (2.20) is dense in the class of valid angular measures (Fougères et al. 2013). In other words, any extremal dependence structure can be arbitrarily well-approximated by that of a max-linear model with sufficiently many factors. This makes max-linear modelling a versatile and powerful framework, despite its simplicity.

There are several ways to construct a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with angular measure (2.20). This thesis uses two constructions. Let $Z_1, \ldots, Z_q$ be independent Fréchet random variables with unit scale and shape parameter α, and set $\boldsymbol{Z} = (Z_1, \ldots, Z_q)$. The two constructions are

$$\boldsymbol{X} = A \times_{\max} \boldsymbol{Z} := \left(\bigvee_{j=1}^q a_{1j} Z_j, \ldots, \bigvee_{j=1}^q a_{dj} Z_j \right) \tag{2.21}$$

and

$$\boldsymbol{X} = A \otimes \boldsymbol{Z} := \bigoplus_{j=1}^{q} (\boldsymbol{a}_j \odot Z_j). \tag{2.22}$$

Adopting the terminology of Cooley and Thibaud (2019), we refer to these as the max-stable and transformed-linear constructions, respectively. Under the max-stable construction, each component X_i is the maximum of linear combinations of the heavy-tailed latent variables $Z_1, \ldots, Z_q$. The second construction, employed in Cooley and Thibaud (2019), is defined in terms of vector space operations $\oplus$ and $\odot$ defined therein. These operations will be defined explicitly and discussed later in Section XX. The difference between the two constructions manifests in their realisations, as illustrated in Figure 7 in the Supplementary Material of Cooley and Thibaud (2019). The directions of large realisations of the max-stable construction tend to correspond almost exactly to the points $\boldsymbol{a}_1/\|\boldsymbol{a}_1\|_\alpha, \ldots, \boldsymbol{a}_q/\|\boldsymbol{a}_q\|_\alpha$. Under the transformed-linear construction, the directions of extreme events tend to lie in a neighbourhood of, but not exactly on, these discrete locations.

Computing joint tail event probabilities is straightforward under the max-linear model. Suppose $\boldsymbol{X}$ is max-linear with parameter matrix A. Consider the extreme failure region

$$\mathcal{R}_f(x) := \{\boldsymbol{y} \in \mathbb{R}_+^d : f(\boldsymbol{y}) > x\}$$

for some function $f : \mathbb{R}_+^d \to \mathbb{R}$. Provided the failure region is sufficiently extreme (distant from the origin), then

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{R}_f(x)) \approx \sum_{j=1}^{q} \frac{\|\boldsymbol{a}_j\|_\alpha^\alpha}{r_\star(\boldsymbol{a}_j/\|\boldsymbol{a}_j\|_\alpha)^\alpha}, \tag{2.23}$$

where $r_\star = r_\star(\boldsymbol{\theta})$ is such that $f(r_\star \boldsymbol{\theta}) = x$ (Cooley and Thibaud 2019; Kiriliouk and Zhou 2022). The formulae corresponding to some popular failure regions are listed below:

$$\begin{aligned}
f(\boldsymbol{y}) &= \max \boldsymbol{y}, & \mathbb{P}(\max \boldsymbol{X} > x) &\approx \sum_{j=1}^{q} \max_{i=1,\ldots,d} \frac{a_{ij}}{x} \\
f(\boldsymbol{y}) &= \min \boldsymbol{y}, & \mathbb{P}(\min \boldsymbol{X} > x) &\approx \sum_{j=1}^{q} \min_{i=1,\ldots,d} \frac{a_{ij}}{x} \\
f(\boldsymbol{y}) &= \boldsymbol{v}^T \boldsymbol{y}, & \mathbb{P}(\boldsymbol{v}^T \boldsymbol{X} > x) &\approx \sum_{j=1}^{q} \frac{\boldsymbol{v}^T \boldsymbol{a}_j}{x}.
\end{aligned}$$

The first and second regions concern extreme events affecting at least one variable or all variables simultaneously, respectively. For the third region, the weight vector $\boldsymbol{v}$ satisfies $v_i \geq 0$ and $v_1 + \ldots + v_d = 1$. Such regions are of interest for climate event attribution (Kiriliouk and Naveau 2020) or quantifying the Value-at-Risk of an asset portfolio (Yuen and Stoev 2014b). Each of these failure probabilities may be perceived as a measure of risk. Risk mitigation is the practice of taking action – bolstering flood defences or diversifying a portfolio – to ensure these probabilities are acceptably small.

2.2.5 Multivariate regular variation

Multivariate regular variation (MRV) provides an alternative framework for characterising the probabilistic structure of the joint tail of random vectors. By imposing a regularity structure on the joint tail, MRV facilitates the development of theoretically justified procedures for extrapolating the probability law from moderately large values to more extreme tail regions. We introduce the concept of regular variation in the univariate setting before extending to the multivariate case.

Definition 2.6. A function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying with index $\alpha \in \mathbb{R}$ if, for all $x > 0$,

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^{\alpha}. \tag{2.24}$$

If $\alpha = 0$, then f is called slowly-varying. Intuitively, a regularly varying function is one that behaves like a power function as the argument approaches infinity. This notion is generalised to random variables by taking the distributional tail as the function of interest.

Definition 2.7. A non-negative random variable X is regularly varying with tail index $\alpha \geq 0$ if the right-tail of its distribution function is regularly varying with index $-\alpha$, i.e. for all $x > 1$,

$$\lim_{t\to\infty} \mathbb{P}(X > tx \mid X > t) = x^{-\alpha}.$$

If X is regularly varying with index α, then its survivor function is of the form

$$\mathbb{P}(X > x) = x^{-\alpha} L(x) \tag{2.25}$$

for some slowly-varying function L (Jessen and Mikosch 2006). Regularly varying random variables are those with power law tails. In fact, a random variable X is regularly varying if and only if it belongs to the Fréchet MDA (CITE). Crucially, (2.25) reveals that regularly varying distributions possess asymptotic scale invariance, in the sense that for all $\lambda > 0$,

$$\mathbb{P}(X > \lambda x) = (\lambda x)^{-\alpha} L(\lambda x) \sim \lambda^{-\alpha} \mathbb{P}(X > x).$$

The ubiquity of regular variation in extreme value statistics is due to this homogeneity property. Under regular variation, the probability law of X at some level λx is identical to the probability law at level λ, up to some constant factor. An analogous interpretation holds when regular variation is generalised to multivariate random vectors, where the joint tail distribution is represented by a homogeneous limit measure.

Although MRV can be formulated more generally – see Section 6.5.5 in Resnick (2007) – we exclusively focus on random vectors $\boldsymbol{X}$ taking values on the positive orthant $\mathbb{R}^d_+ := [0, \infty)^d$. This common assumption is not as restrictive as it might initially seem. In most applications, the risk being assessed is directional. For example, a climatologist might model the lows or the highs of precipitation records depending on they are analysing drought risk or flood risk. Without loss of generality and by means of a transformation if necessary, this direction of interest can be defined as 'positive'.

Definition 2.8. A random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ is multivariate regularly varying with tail index $\alpha > 0$, denoted $\boldsymbol{X} \in \mathcal{RV}^d_+(\alpha)$, if it satisfies the following (equivalent) statements (Resnick 2007):

1. There exists a sequence $b_n \to \infty$ and a non-negative Radon measure ν on $\mathbb{E}_0 := [0, \infty]^d \setminus \{\mathbf{0}\}$ such that

$$n\mathbb{P}(b_n^{-1} \boldsymbol{X} \in \cdot) \xrightarrow{\text{v}} \nu(\cdot), \qquad (n \to \infty), \tag{2.26}$$

where $\xrightarrow{\text{v}}$ denotes vague convergence in the space of non-negative Radon measures on $\mathbb{E}_0$. The exponent measure ν is homogeneous of order $-\alpha$, that is, for any $s > 0$,

$$\nu(s\,\cdot) = s^{-\alpha} \nu(\cdot). \tag{2.27}$$

2. Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^d$. Denote the radial and angular components of $\boldsymbol{X}$ by $R := \|\boldsymbol{X}\|$ and $\boldsymbol{\Theta} := \boldsymbol{X}/\|\boldsymbol{X}\|$. Then there exists a sequence $b_n \to \infty$ and a finite measure H on the simplex

$$\mathbb{S}_+^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\| = 1\} \tag{2.28}$$

such that

$$n\mathbb{P}((b_n^{-1}R, \boldsymbol{\Theta}) \in \cdot) \xrightarrow{\text{v}} \nu_\alpha \times H(\cdot), \qquad (n \to \infty), \tag{2.29}$$

in the space of non-negative Radon measures on $(0, \infty] \times \mathbb{S}_+^{d-1}$, where $\nu_\alpha((x, \infty)) = x^{-\alpha}$ for any $x > 0$.

The limit measures ν and H in (2.26) and (2.29) are related via

$$\nu(\{\boldsymbol{x} \in \mathbb{E}_0 : \|\boldsymbol{x}\| > s, \boldsymbol{x}/\|\boldsymbol{x}\| \in \cdot\}) = s^{-\alpha}H(\cdot), \qquad \nu(\mathrm{d}r \times \mathrm{d}\boldsymbol{\theta}) = \alpha r^{-\alpha-1}\mathrm{d}r\,\mathrm{d}H(\boldsymbol{\theta}). \tag{2.30}$$

The attractive feature of MRV is best represented by its pseudo-polar formulation (2.29). This states that the extremal behaviour of $\boldsymbol{X}$ is fully characterised by two quantities: the tail index and the angular measure. The tail index α represents the index of regular variation of the (univariate) radial component. It governs the heavy-tailedness of the size (norm) of $\boldsymbol{X}$. The angular measure H fully characterises the dependence structure. Crucially, the right-hand side of (2.29) is a product measure, signifying that the radial and angular components are independent in the limit.

The MRV property implicitly requires that the marginal components $X_1, \ldots, X_d$ are heavy-tailed with a shared tail index. Standard practice is to standardise the margins prior to modelling the dependence structure (Section XX), so this is not restrictive. In this thesis, we will always choose Fréchet margins with unit scale and shape parameter $\alpha > 0$, that is

$$\mathbb{P}(X_i < x) = \exp(-x^{-\alpha}), \qquad (x > 0). \tag{2.31}$$

An MRV random vector on α-Fréchet margins (2.31) has tail index α. Thus, as before, fixing the margins deals with the tail index and the angular measure becomes the object of interest.

The angular measure is unique only with respect to a pre-specified norm $\|\cdot\|$ and lies on

the corresponding unit simplex (2.28). As mentioned previously, we exclusively choose the L_p-norm

$$\|\cdot\|_p : \mathbb{R}^d \to \mathbb{R}, \qquad \|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{d} x_i^p\right)^{1/p} \tag{2.32}$$

with (2.11) the corresponding simplex. The mass of the angular measure is $m := H(\mathbb{S}_+^{d-1}) \in (0, \infty)$. The sequence $\{b_n\}$ and the quantity m are jointly determined by (2.29). Replacing $\{b_n\}$ by $\{sb_n\}$ for some $s > 0$ yields a new angular measure $H' = s^{-\alpha}H$ whose mass is $m' = s^{-\alpha}m$. We are free to choose whether the scaling information is contained in $\{b_n\}$ or m. Possible reasons for preferring one over the other are discussed in Fougères et al. (2013), but ultimately it is an arbitrary modelling choice. In previous sections, H was normalised to be a probability measure with $m = 1$. Henceforth, we will tend to specify $\{b_n\}$ and push the scaling information on to H. With $\boldsymbol{X}$ standardised to α-Fréchet margins, the centre of mass of H must lie in the simplex interior:

$$\int_{\mathbb{S}_+^{d-1}} \theta_i \, \mathrm{d}H(\boldsymbol{\theta}) = \mu > 0, \qquad (i = 1, \ldots, d). \tag{2.33}$$

Were this not the case it would imply that at least one variable can never be extreme, contradicting the assumption that all variables have equally heavy tails. The value of μ depends on the choice of norm and the mass of H. If $\|\cdot\| = \|\cdot\|_1$, then $\mu = m/d$ in accordance with (2.12). If $\|\cdot\| = \|\cdot\|_2$, then $m/d \leq \mu \leq m/\sqrt{d}$ according to Lemma 2.1 in Fomichov and Ivanovs (2023). The lower and upper bounds are attained when H places all its mass at the vertices of the simplex or at its centre, respectively. These can be understood as the limiting cases of extremal dependence, which is formalised in the next section.

2.2.6 Extremal dependence measures

The extremal dependence structure of a random vector $\boldsymbol{X}$ can be quantified and classified using a plethora of summary measures (Coles, Heffernan, et al. 1999). We focus on the tail dependence coefficient and the extremal dependence measure.

2.2.6.1 The tail dependence coefficient

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 bulk of the distribution but exhibit dependence in their extremes, or vice versa. The extremal dependence structure may be very complex; angular measures form an infinite-dimensional class subject only to a set of moment constraints. For example, suppose X_i and X_j represent the recorded values of a meteorological variable measured at two spatial locations. The extremal dependence between X_i and X_j may depend on the spatial proximity of the sites, the topography of the spatial domain, the physics of the climatological process, and a multitude of other factors. The complexity grows as more variables are introduced, as higher-order dependencies come into play. Extremal dependence measures aim to provide summary information about particular aspects of the dependence structure. One such measure is the tail dependence coefficient (CITE).

Definition 2.9. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ with $X_i \sim F_i$ for $i = 1, \ldots, d$. Let $\beta \subseteq \{1, \ldots, d\}$ with $|\beta| \geq 2$ and define $\boldsymbol{X}_\beta := \{X_i : i \in \beta\}$. The tail dependence coefficient associated with β is (CITE e.g. Simpson et al 2020)

$$\chi_\beta = \lim_{u \to 1} \chi_\beta(u) = \lim_{u \to 1} \frac{\mathbb{P}(F_i(X_i) > u : i \in \beta)}{1 - u}. \tag{2.34}$$

When $\beta = \{i, j\}$ for $i \neq j$, we write $\chi_\beta =: \chi_{ij}$.

We say that X_i and X_j are asymptotically independent (AI) if and only if $\chi_{ij} = 0$. Asymptotic independence means that both variables cannot take extreme values simultaneously. If $\chi_{ij} \in (0, 1]$, then the variables are asymptotically dependent (AD) and may be simultaneously extreme. The interpretation of χ_β for $|\beta| > 2$ is more subtle. If $\chi_\beta \in (0, 1]$, then all components of $\boldsymbol{X}_\beta$ may be simultaneously large. If $\chi_\beta = 0$, then the corresponding variables may not be concomitantly extreme, but this does not preclude the possibility that $\chi_{\beta'} > 0$ for some $\beta' \subset \beta$ with $|\beta'| \geq 2$.

The nullity of otherwise of the tail dependence coefficients is determined by which subspaces of the simplex are charged with H-mass. Specifically, $\chi_\beta > 0$ if and only if there exists

$\beta' \supseteq \beta$ such that

$$H(\{\boldsymbol{\theta} \in \mathbb{S}_+^{d-1} : \theta_i > 0 \iff i \in \beta'\}) > 0. \tag{2.35}$$

For example, consider the angular measures

$$H^{(1)} = \frac{m}{d} \sum_{i=1}^{d} \delta_{\boldsymbol{e}_i}, \qquad H^{(2)} = m\delta_{\mathbf{1}/\|\mathbf{1}\|}, \tag{2.36}$$

where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$ denote the canonical basis vectors of $\mathbb{R}^d$. The measure $H^{(1)}$ places all its mass on the vertices of the simplex. This corresponds to full asymptotic independence, since then $\chi_\beta = 0$ for all $\beta \subseteq \{1, \ldots, d\}$ with cardinality at least equal to two. The angular measure $H^{(2)}$ concentrates at a single point at the centre of the simplex. This implies that $\chi_{\{1,\ldots,d\}} > 0$ and consequently $\chi_\beta > 0$ for all subsets β.

If the bivariate exponent measure V_{ij} of (X_i, X_j) is known, then the tail dependence coefficient χ_{ij} may be computed using the relation $\chi_{ij} = 2 - V_{ij}(1,1)$ (Coles, Heffernan, et al. 1999). The following examples illustrate this for selected parametric models.

Example 2.1. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be symmetric logistic distributed with dependence parameter $\gamma \in (0, 1]$. For any $i \neq j$, let V_{ij} denote the bivariate exponent measure of (X_i, X_j). Then

$$\chi_{ij} = 2 - V_{ij}(1,1) = 2 - \left[\left(x_i^{-1/\gamma} + x_j^{-1/\gamma}\right)^{\gamma}\right] = 2 - 2^{\gamma}.$$

Therefore X_i and X_j are asymptotically independent when $\gamma = 1$ and approach complete asymptotic dependence as $\gamma \to 0$.

Example 2.2. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be Hüsler-Reiss distributed with parameter matrix $\Lambda = (\lambda_{ij}^2)$. For any $i \neq j$, let V_{ij} denote the bivariate exponent measure of (X_i, X_j). Then

$$\chi_{ij} = 2 - V_{ij}(1,1) = 2 - 2\Phi\left(\lambda_{ij} + \frac{1}{2\lambda_{ij}} \log 1\right) = 2 - 2\Phi(\lambda_{ij}),$$

where Φ is the standard normal distribution function. Variables X_i and X_j are asymptotically dependent for all $\lambda_{ij} > 0$, with asymptotic independence in the limit as $\lambda_{ij} \to \infty$. *Refer back to this equation when discussing Hazra and Bose changepoint method – it gives*

one-to-one relationship between HR parameter and dependence strength, so testing for change in λ or χ are equivalent.

Example 2.3. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d)$ is max-linear with parameter matrix $A \in \mathbb{R}_+^{d\times q}$. Substituting (2.20) into (2.10) yields

$$\chi_{ij} = 2 - V_{12}(1,1) = 2 - 2\int_{\mathbb{S}^1_{+(1)}} (\theta_1 \vee \theta_2)\,\mathrm{d}H(\boldsymbol{\theta}) = 2 - \sum_{l=1}^{q}(a_{il} \vee a_{jl}). \tag{2.37}$$

Consider two max-linear random vectors with discrete angular measures $H^{(1)}$ and $H^{(2)}$ as in (2.36). The parameter matrices are given by

$$A^{(1)} = I_d \in \mathbb{R}_+^{d\times d}, \qquad A^{(2)} = \mathbf{1}_d \in \mathbb{R}_+^{d\times 1}.$$

The tail dependence coefficients under these models are

$$\chi_{ij}^{(1)} = 2 - \sum_{j=1}^{2}\max(0,1) = 0, \qquad \chi_{ij}^{(2)} = 2 - \sum_{j=1}^{1}\max(1,1) = 1,$$

corresponding to complete dependence and asymptotic dependence, as expected.

Estimates of χ_{ij} are obtained by estimating $\hat{\chi}_{ij}(u)$ at a sequence of high quantiles u approaching one. The `taildep` function in the R package `extRemes` achieves this using the estimator given in Equation 2.62 in Reiss and Thomas (2007) and produces a diagnostic plot as shown in Figure 2.1. For this example the data were generated from a symmetric logistic model with $\gamma = 0.5$. The horizontal dashed line indicates the true value $\chi_{ij} = 2 - \sqrt{2} \approx 0.59$, while the blue points represent the estimates $\hat{\chi}_{ij}(u)$ over the range $0.8 \leq u \leq 0.995$. The shaded region depicts the 95% Wald confidence interval. We encounter a bias-variance trade-off in relation to quantile/threshold, similar in nature to that described in Section XX with respect to the selecting the block size/threshold.

Estimation of χ_β for $|\beta| > 2$ is more complicated and is related to the task of determining the support of the angular measure (Goix et al. 2017; Meyer and Wintenberger 2023; Simpson et al. 2020). This thesis primarily concerns dependence at the pairwise level, so we direct the reader to the aforementioned papers and the review Engelke and Jevgenijs Ivanovs (2021) for further details.

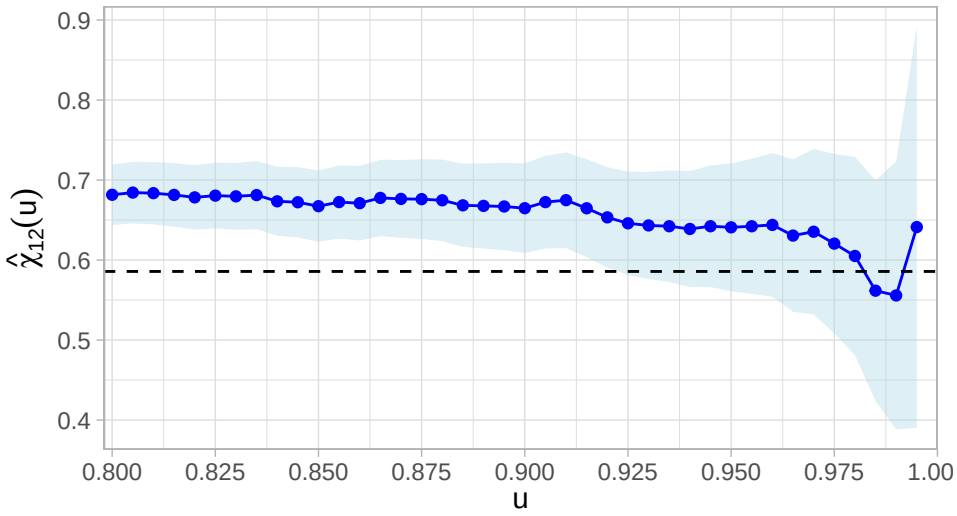

Figure 2.1: Empirical estimates $\hat{\chi}_{12}(u)$ of the tail dependence coefficient for bivariate symmetric logistic data with $\gamma = 0.5$ and $n = 5,000$ observations. The true coefficient $\chi_{12} = 2 - 2^{\gamma} \approx 0.59$ is marked by the dashed line. The shaded region represents the 95% Wald confidence interval.

Let $\chi = (\chi_{ij})$ denote the Tail Dependence Matrix (TDM) of bivariate tail dependence coefficients with diagonal entries $\chi_{ii} := 1$. The TDM provides a high level summary of the extremal dependence structure. It has been applied for exploratory analysis (Huang et al. 2019) and considered as a tool for clustering (Fomichov and Ivanovs 2023). Other works focus on its theoretical properties. Shyamalkumar and Tao (2020) conjecture that the 'realisation problem' – determining whether a given matrix is a valid TDM – is NP-complete; this was recently proved by Janßen, Neblung, et al. (2023). By establishing a correspondence between the class of TDMs and a metric space, Janßen, Neblung, et al. (2023) also show that, in certain cases, higher order tail-dependence is determined by the bivariate TDM. Section XX introduces a similar (and similarly named) matrix, the Tail *Pairwise* Dependence Matrix (TPDM), which is the eponym of this thesis. Rather than the tail dependence coefficient χ_{ij}, the TPDM is founded on an alternative bivariate summary measure called the Extremal Dependence Measure (EDM).

2.2.6.2 Extremal dependence measure

The extremal dependence measure (EDM) is a pairwise summary measure similar to χ_{ij}. It was originally proposed Resnick (2004) and later generalised by Larsson and Resnick (2012).

Definition 2.10. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ with angular measure H. The EDM between X_i and X_j is

$$\text{EDM}_{ij} := \int_{\mathbb{S}_+^{d-1}} \theta_i \theta_j \, \mathrm{d}H(\boldsymbol{\theta}). \tag{2.38}$$

The EDM depends on the choice of norm via the angular measure, but Larsson and Resnick (2012) show that EDMs under different norms are equivalent in a certain sense. The EDM was originally defined by Resnick (2004) for bivariate random vectors $\boldsymbol{X} = (X_1, X_2)$. In their definition, the integrand is

$$\left(\frac{4}{\pi}\right)^2 \arctan\left(\frac{\theta_2}{\theta_1}\right)\left[\frac{\pi}{2} - \arctan\left(\frac{\theta_2}{\theta_1}\right)\right]. \tag{2.39}$$

rather than $\theta_1\theta_2$. The original and refined versions are also equivalent.

Being explicitly defined in terms of the angular measure, the EDM's interpretation in terms of AD/AI is straightforward. Recall from (2.35) that variables X_i and X_j are asymptotically independent if and only if $H(\{\boldsymbol{\theta} : \theta_i, \theta_j > 0\}) = 0$. Then

$$\chi_{ij} = 0 \iff \int_{\{\boldsymbol{\theta} \in \mathbb{S}_+^{d-1} \,:\, \theta_i, \theta_j > 0\}} \theta_i \theta_j \, \mathrm{d}H(\boldsymbol{\theta}) = 0 \iff \text{EDM}_{ij} = 0.$$

The EDM is maximal when X_i and X_j are perfectly asymptotically dependent. The maximal value depends on the choice of norm and the mass of the angular measure. When $d = 2$ and $\|\cdot\| = \|\cdot\|_p$ we have $\text{EDM}_{ij} \leq 2^{-2/p} m$ with equality if and only if H places all its mass at the simplex barycentre, that is $H(\{(2^{-1/p}, 2^{-1/p})\}) = m$.

We return to the EDM in Section XX when introducing the tail pairwise dependence matrix.

2.3 Inference

We now shift our attention to the topic of (non-parametric) inference in multivariate extremes. The general approach entails using the angular components of large observations to learn a model for H. This strategy is justified by the MRV assumption: (2.29) implies that

$$\boldsymbol{\Theta} \mid (R > t) \xrightarrow{d} H(\cdot), \qquad (t \to \infty). \tag{2.40}$$

The angular measure is the limiting distribution of the angles of exceedances of some radial threshold. By analogy to the peaks-over-threshold approach (Section XX), it suggests itself to base inference on the subset of data points whose norm exceeds some high fixed threshold. Increasing the threshold reduces the number of observations that enter into the estimators, and vice versa. It is generally more convenient to specify the desired number of threshold exceedances, denoted k, and set the threshold accordingly. This approach is most conveniently described using order statistics.

2.3.1 Framework and notation

Consider a d-dimensional MRV random vector $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ denote a sequence of independent copies of $\boldsymbol{X}$ and fix a norm $\|\cdot\|$ on $\mathbb{R}^d$. For $i \geq 1$, denote by

$$R_i := \|\boldsymbol{X}_i\|, \qquad \boldsymbol{\Theta}_i := (\Theta_{i1}, \ldots, \Theta_{id}) = \frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|}, \tag{2.41}$$

the radial and angular components of $\boldsymbol{X}_i$ with respect to the chosen norm. Assume that the distribution of $\|\boldsymbol{X}\|$ is continuous. Then for any $n \geq 1$, there exists a permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that

$$\|\boldsymbol{X}_{(1),n}\| > \|\boldsymbol{X}_{(2),n}\| > \ldots > \|\boldsymbol{X}_{(n),n}\|,$$

where $\boldsymbol{X}_{(i),n} := \boldsymbol{X}_{\pi(i)}$ for $i = 1, \ldots, n$. The random variable $\|\boldsymbol{X}_{(j),n}\|$ is called the jth (upper) order statistic of $\{\|\boldsymbol{X}_i\| : i = 1, \ldots, n\}$. Henceforth, we suppress the dependence on n in our order statistic notation. Let the radial and angular components of $\boldsymbol{X}_{(i)}$ be denoted by

$$R_{(i)} = \|\boldsymbol{X}_{(i)}\|, \qquad \boldsymbol{\Theta}_{(i)} = (\Theta_{(i),1}, \ldots, \Theta_{(i),d}) = \frac{\boldsymbol{X}_{(i)}}{\|\boldsymbol{X}_{(i)}\|}. \tag{2.42}$$

Performing inference based on the $k = k(n)$ largest observations is equivalent to performing inference based on the set of observations whose norm exceeds the threshold $t = R_{(k+1)}$.

2.3.2 Selecting the radial threshold or the number of exceedances

All estimators will require on choosing the number of extreme observations k that enter into them. In theoretical analyses, it is customary to choose the sequence $\{k(n) : n \geq 1\}$

such that

$$\lim_{n\to\infty} k(n) = \infty, \qquad \lim_{n\to\infty} \frac{k(n)}{n} = 0. \tag{2.43}$$

These arise as sufficient conditions for proving various asymptotic properties (e.g. consistency, asymptotic normality) of estimators. The condition $k \to \infty$ ensures that the number of extremes – the effective sample size – grows arbitrarily large. The second condition $k/n \to 0$ requires that the proportion of threshold exceedances becomes vanishingly small, ensuring that inference is targeting the tail. In practice, n is fixed and selecting k requires striking a balance between these two aspects. Choosing k too small reduces the amount of available information and leads to unnecessarily high uncertainty. If k is too large, we risk using data that does not reflect the extremal dependence structure leading to bias. An appropriate choice depends on both the sample size and the underlying distribution of $\boldsymbol{X}$. If the convergence in (2.40) is rapid, then a low threshold may be adequate. Several threshold selection procedures have been proposed in univariate extremes (Section XX), but the literature on radial threshold selection is comparatively scant. By combining two sub-tests regarding (i) independence of the radial and angular components and (ii) regular variation of the radial component, Einmahl, Yang, et al. (2020) devise a formal procedure testing the validity of the MRV assumption. They suggest choosing the threshold by examining a plot of the sequence of p-values against k. The support-detection algorithm of Meyer and Wintenberger (2023) chooses k automatically via minimisation of a penalised log-likelihood. This procedure is specific to their setting and relies on additional technical assumptions. Most applied studies use a rule-of-thumb approach and/or produce a threshold stability plot checking the (in)sensitivity of their results to the choice of k – see Jiang et al. (2020) and Russell and Hogan (2018) for examples.

2.3.3 The empirical angular measure

Once the tuning parameter k has been chosen, attention turns towards the extremal angles $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}$. In view of (2.40), the empirical distribution of $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}$ is the natural non-parametric estimator for the angular measure.

Definition 2.11. The empirical angular measure based on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is the random

measure on $\mathbb{S}_+^{d-1}$ defined as

$$\hat{H}(\cdot) := \frac{m}{k} \sum_{i=1}^{n} \delta_{\boldsymbol{\Theta}_i}(\cdot) \mathbf{1}\{R_i > R_{(k+1)}\} = \frac{m}{k} \sum_{i=1}^{k} \delta_{\boldsymbol{\Theta}_{(i)}}(\cdot). \tag{2.44}$$

Note that $\hat{H}$ does not enforce the moment constraints (2.12), so is not necessarily a valid angular measure. Einmahl and Segers (2009) construct an alternative non-parametric estimator that does enforce these restrictions, but it is limited to the bivariate setting. Proposition 3.3 in Janßen and Wan (2020) establishes consistency $\hat{H} \xrightarrow{p} H$ of the empirical angular measure provided the level k satisfies the rate conditions (2.43). Their result holds for general norms in arbitrary dimensions. Clémençon et al. (2023) conduct a non-asymptotic (i.e. finite sample) analysis of $\hat{H}$, establishing high-probability bounds on the worst-case estimation error $\sup_{A \in \mathcal{A}} |H(A) - \hat{H}(A)|$ over classes $\mathcal{A}$ of Borel subsets on $\mathbb{S}_+^{d-1}$. Their results hold with $\|\cdot\| = \|\cdot\|_p$ for $p \in [1, \infty]$. Since $\hat{H}$ is a discrete measure concentrating at k points, there exists a max-linear random vector $\boldsymbol{X}$ with parameter matrix

$$\hat{A} := \left(\frac{m}{k}\right)^{1/\alpha} \left(\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}\right) \in \mathbb{R}_+^{d \times k}. \tag{2.45}$$

whose angular measure is $\hat{H}$. Estimates of tail event probabilities under the empirical model $\hat{H}$ may then be computed using the formula (2.23).

2.3.4 Non-parametric estimators

Larsson and Resnick (2012) remark that analysing extremal dependence often involves quantities of the form

$$\mathbb{E}_H[f(\boldsymbol{\Theta})] := \int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta}) \, \mathrm{d}H(\boldsymbol{\theta}) = \mathbb{E}_{m^{-1}H}[m f(\boldsymbol{\Theta})], \tag{2.46}$$

where $f : \mathbb{S}_+^{d-1} \to \mathbb{R}$. We have already seen an example of this in Definition 2.10: the EDM between X_i and X_j is defined as (2.46) with $f(\boldsymbol{\theta}) = \theta_i \theta_j$. We reiterate that in our notation, the expectation is with respect to a measure H that is not necessarily normalised. When manipulating expectations/variances, the following relations may be useful to bear

in mind:

$$\mathbb{E}_H[f(\boldsymbol{\Theta})] = \mathbb{E}_{m^{-1}H}[mf(\boldsymbol{\Theta})] = m\mathbb{E}_{m^{-1}H}[f(\boldsymbol{\Theta})]$$
$$\mathrm{Var}_H[f(\boldsymbol{\Theta})] = \mathbb{E}_{m^{-1}H}[m^2 f(\boldsymbol{\Theta})^2] - \mathbb{E}_{m^{-1}H}[mf(\boldsymbol{\Theta})]^2 = m^2\mathrm{Var}_{m^{-1}H}[f(\boldsymbol{\Theta})].$$

Klüppelberg and Krali (2021) opt to normalise H and absorb m into f. For example, the EDM would correspond to $f(\boldsymbol{\theta}) = m\theta_i\theta_j$ in their notation. Suppressing the normalising constant arguably results in less cumbersome notation, but in any case the choice is purely stylistic.

To construct non-parametric estimators of quantities (2.46), we simply replace H with the empirical angular measure $\hat{H}$, yielding (Klüppelberg and Krali 2021)

$$\hat{\mathbb{E}}_H[f(\boldsymbol{\Theta})] := \mathbb{E}_{\hat{H}}[f(\boldsymbol{\Theta})] = \int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta})\,\mathrm{d}\hat{H}(\boldsymbol{\theta}) = \frac{m}{k}\sum_{i=1}^{k} f(\boldsymbol{\Theta}_{(i)}). \tag{2.47}$$

Klüppelberg and Krali (2021) prove asymptotic normality of these estimators by generalising a result in Larsson and Resnick (2012).

Theorem 2.3. *Let $f : \mathbb{S}_+^{d-1} \to \mathbb{R}$ be continuous and assume k satisfies the rate conditions (2.43). Moreover, suppose that*

$$\lim_{n\to\infty} \sqrt{k}\left[\frac{n}{k}\mathbb{E}[f(\boldsymbol{\Theta}_1)\mathbf{1}\{R_1 \geq b_{\lfloor n/k\rfloor}t^{-1/\alpha}\}] - \mathbb{E}_H[f(\boldsymbol{\Theta})]\frac{n}{k}\bar{F}_R(b_{\lfloor n/k\rfloor}t^{-1/\alpha})\right] = 0 \tag{2.48}$$

holds locally uniformly for $t \in [0,\infty)$, where $\bar{F}_R(\cdot) = \mathbb{P}(R > \cdot)$ denotes the survivor function of R. Finally, assume that

$$\nu^2 := \mathrm{Var}_H(f(\boldsymbol{\Theta})) > 0. \tag{2.49}$$

Then

$$\sqrt{k}\left[\hat{\mathbb{E}}_H[f(\boldsymbol{\Theta})] - \mathbb{E}_H[f(\boldsymbol{\Theta})]\right] \to N(0,\nu^2), \qquad (n\to\infty). \tag{2.50}$$

The rate condition (2.48) requires that the dependence between the radius and angle decays sufficiently quickly. This condition is non-observable and must be assumed.

Example 2.4. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent copies of $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$. The estimator for the EDM between X_i and X_j is

$$\widehat{\mathrm{EDM}}_{ij} := \hat{\mathbb{E}}_H[\Theta_i \Theta_j] \frac{m}{k} \sum_{l=1}^{k} \Theta_{(l),i} \Theta_{(l),j}.$$

Under the conditions of Theorem 2.3,

$$\sqrt{k}[\widehat{\mathrm{EDM}}_{ij} - \mathrm{EDM}_{ij}] \to N(0, \nu_{ij}^2), \qquad \nu_{ij}^2 = \mathrm{Var}_H(\Theta_i \Theta_j).$$

2.4 Tail pairwise dependence matrix (TPDM)

This section introduces the key protagonist of this thesis: the tail pairwise dependence matrix (TPDM).

2.4.1 Definition and examples

Preamble.

Definition 2.12. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(2)$ with normalising sequence $b_n = n^{1/2}$. Let H denote the angular measure with respect to $\|\cdot\|_2$. The TPDM of $\boldsymbol{X}$ is the $d \times d$ matrix

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} = \int_{\mathbb{S}_{+(2)}^{d-1}} \theta_i \theta_j \, \mathrm{d}H(\boldsymbol{\theta}) = \mathbb{E}_H[\Theta_i \Theta_j]. \tag{2.51}$$

The TPDM is essentially a matrix of EDMs subject to additional restrictions on the tail index, normalising sequence, and norm. Each off-diagonal entry σ_{ij} may be interpreted as summarising the dependence between X_i and X_j, with $\sigma_{ij} = 0$ if and only if the corresponding variables are asymptotically independent. The original definition was generalised by Kiriliouk and Zhou (2022) to permit general α.

Definition 2.13. For $\alpha \geq 1$, let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ with normalising sequence $b_n = n^{1/\alpha}$. Let H denote the angular measure with respect to $\|\cdot\|_\alpha$. The TPDM of $\boldsymbol{X}$ is the $d \times d$ matrix

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} = \int_{\mathbb{S}_{+(\alpha)}^{d-1}} \theta_i^{\alpha/2} \theta_j^{\alpha/2} \, \mathrm{d}H(\boldsymbol{\theta}) = \mathbb{E}_H[\Theta_i^{\alpha/2} \Theta_j^{\alpha/2}]. \tag{2.52}$$

The tail index of $\boldsymbol{X}$ is now arbitrary, but the normalisation sequence and norm are still required to conform with this index. It is obvious that these definitions coincide when $\alpha = 2$, but Kiriliouk and Zhou (2022) provide no direct rationale for why (2.52) is the natural generalisation of (2.51). Appendix XX provides a series of results shedding light on this matter. After generalising a result in Fix et al. (2021) (Lemma A.1), we prove that the TPDM is invariant to the choice of α (Proposition A.1). This culminates in an expression for the TPDM (for any α) in terms of the L_1 angular density that does not depend on α. We now use of this formula and the angular densities in Semadeni (2020) to compute the TPDM under the symmetric logistic and Hüsler-Reiss models. These model TPDMs will be especially useful in Chapter XX for evaluating the performance of TPDM estimators.

Example 2.5. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ follows the symmetric logistic distribution with dependence parameter $\gamma \in (0, 1)$. For any $i \neq j$,

$$\sigma_{ij} = \frac{1-\gamma}{\gamma} \int_0^1 [u(1-u)]^{\frac{1}{\gamma}-\frac{3}{2}} [(1-u)^{1/\gamma} + u^{1/\gamma}]^{\gamma-2} \, \mathrm{d}u. \tag{2.53}$$

Example 2.6. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ follows the Hüsler-Reiss distribution with parameter matrix $\Lambda = (\lambda_{ij}^2)$. For any $i \neq j$,

$$\sigma_{ij} = \int_0^1 \frac{\exp(-\lambda_{ij}/4)}{2\lambda_{ij} u(1-u)} \phi\left(\frac{1}{2\lambda_{ij}} \log\left(\frac{u}{1-u}\right)\right) \mathrm{d}u. \tag{2.54}$$

The blue lines in Figure 2.2 plot (2.53) and (2.54) against the model parameter. For comparison, we also include the tail dependence coefficients (red lines) computed using Example 2.1 and Example 2.2. For both models, the strength of association is a decreasing function of the model parameter, with complete dependence (resp. asymptotic independence) as the parameter approaches zero (resp. its upper limit). For the Hüsler-Reiss distribution, dependence is very weak beyond $\lambda \approx 3$. We can check that this is correct by comparing with Figure 1 in the Supplementary Material of Cooley and Thibaud (2019). The figure reveals that for a Brown-Resnick process with semi-variogram (2.18) with range $\rho = 2.4$ and smoothness $\kappa = 1.8$, dependence vanishes beyond a distance of approximately 12 units. Recall from Section XX that the dependence between two sites h

units apart under the Brown-Resnick model is equivalent to the dependence between two Hüsler-Reiss variables with dependence parameter $\lambda_{ij} = \sqrt{2(h/\rho)^\kappa}/2$. Setting $h = 12$ gives $\lambda_{ij} = \sqrt{2(12/2.4)^{1.8}}/2 \approx 3.01$, corroborating the results of Figure 2.2. Further verification of our expressions are provided by the shaded regions in Figure 2.2. These represent the minimum/maximum values of 10 estimates of χ_{ij} and σ_{ij} for a sequence of values of γ and λ. The estimates are obtained from large samples ($n = 5\times10^5$) so it is reasonable to neglect the influence of estimation error. The empirical estimates agree with our calculations.

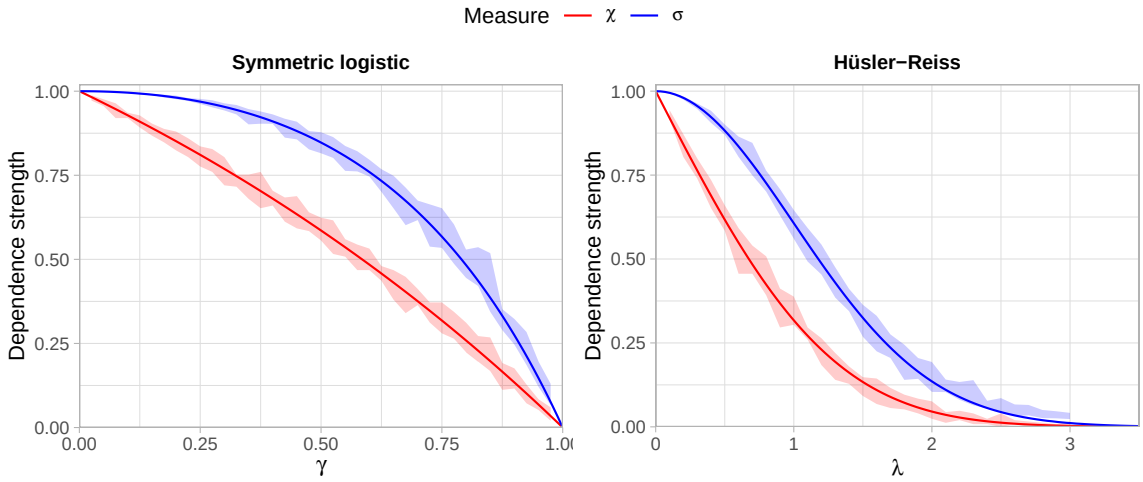

Figure 2.2: True dependence strengths for the symmetric logistic (left) and Hüsler-Reiss (right) models, measured using the tail dependence coefficient (red line) and TPDM (blue line). The shaded regions represent the minimum/maximum values of empirical estimates over 10 repeated simulations using bivariate samples of size $n = 5 \times 10^5$.

The angular measure of a max-linear random vector is discrete, so the angular density does not exist. Nevertheless, it is straightforward to compute the model TPDM directly from the definition (Cooley and Thibaud 2019; Kiriliouk and Zhou 2022).

Example 2.7. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ is max-linear with parameter matrix A. Then for any $i \neq j$,

$$\begin{aligned}
\sigma_{ij} &= \int_{\mathbb{S}_{+(\alpha)}^{d-1}} \theta_i^{\alpha/2}\theta_j^{\alpha/2}\,\mathrm{d}H(\boldsymbol{\theta}) \\
&= \sum_{l=1}^{q} \|\boldsymbol{a}_l\|_\alpha^\alpha \left(\frac{a_{li}}{\|\boldsymbol{a}_l\|_\alpha}\right)^{\alpha/2}\left(\frac{a_{lj}}{\|\boldsymbol{a}_l\|_\alpha}\right)^{\alpha/2} \\
&= \sum_{l=1}^{q} a_{il}^{\alpha/2}a_{jl}^{\alpha/2}.
\end{aligned}$$

Therefore $\Sigma = A^{\alpha/2}(A^{\alpha/2})^T$. Taking A to be $A^{(1)}$ and $A^{(2)}$ as defined in Example 2.3, the corresponding TPDMs are

$$\Sigma^{(1)} = I_d I_d^T = I_d, \qquad \Sigma^{(2)} = \mathbf{1}\mathbf{1}^T = J_d,$$

where J_d is the $d \times d$ all-ones matrix. By construction, these represent the TPDMs under asymptotic dependence and complete dependence, respectively.

The connection between A and Σ will play a prominent role in this thesis. *Say more about this?*

2.4.2 Interpretation of the TPDM entries

The definition of the TPDM

$$\Sigma = \mathbb{E}_H \left[\boldsymbol{\Theta}^{\alpha/2} (\boldsymbol{\Theta}^{\alpha/2})^T \right], \tag{2.55}$$

bears a striking resemblance to the definition of a covariance matrix in the non-extreme setting. The covariance matrix represents the second-order (central) moment of a random vector. Its diagonal entries convey the scale (variance) of the components, while the off-diagonal entries summarise the strength of association (unnormalised correlation) between all pairs of variables. The TPDM entries offer analogous interpretations, except the notions of scale and association are adapted to refer to properties of the joint distributional tail.

Definition 2.14. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ with normalisation sequence b_n. For $i = 1, \ldots, d$, the scale of X_i is defined as (Klüppelberg and Krali 2021)

$$\text{scale}(X_i) = \left[\int_{\mathbb{S}_+^{d-1}} \theta_i^\alpha \, \mathrm{d}H(\boldsymbol{\theta}) \right]^{1/\alpha}.$$

As discussed earlier, a well-defined notion of scale must fix either the sequence b_n or the mass of the angular measure in advance. In the above definition, the normalisation sequence is fixed and scaling information is contained in H. The scale is so-called because it yields

information about the scale of the marginal distributions. Using (2.30), one can show that

$$\begin{aligned}\lim_{n\to\infty} n\mathbb{P}(b_n^{-1}X_i > x) &= \int_{\mathbb{S}_{+(\alpha)}^{d-1}} \int_{x/\theta_i}^{\infty} \alpha r^{-\alpha-1}\,\mathrm{d}r\,\mathrm{d}H(\boldsymbol{\theta}) \\ &= \int_{\mathbb{S}_{+(\alpha)}^{d-1}} [r^{-\alpha}]_{\infty}^{x/\theta_i}\,\mathrm{d}H(\boldsymbol{\theta}) \\ &= x^{-\alpha}[\mathrm{scale}(X_i)]^{\alpha},\end{aligned}$$

Moreover, it behaves like a measure of scale: for any $c > 0$,

$$\begin{aligned}\mathrm{scale}(cX_i) &= \left[\frac{\lim_{n\to\infty} n\mathbb{P}(b_n^{-1}cX_i > x)}{x^{-\alpha}}\right]^{1/\alpha} \\ &= \left[c^{\alpha}\frac{\lim_{n\to\infty} n\mathbb{P}(b_n^{-1}X_i > x/c)}{(x/c)^{-\alpha}}\right]^{1/\alpha} \\ &= c \cdot \mathrm{scale}(X_i).\end{aligned}$$

Comparing Definition 2.14 against Definition 2.13, the diagonal entries of the TPDM are related to the marginal scales via $\mathrm{scale}(X_i) = \sigma_{ii}^{1/\alpha}$. Consequently, if the marginal distributions are standardised to have unit scales, then all diagonal entries of the TPDM are equal to one. Moreover, when $b_n = n^{1/\alpha}$ and $\|\cdot\| = \|\cdot\|_{\alpha}$, the mass of the angular measure relates to the marginal scales via

$$\sum_{i=1}^{d}\sigma_{ii} = \sum_{i=1}^{d}\int_{\mathbb{S}_{+(\alpha)}^{d-1}}\theta_i^{\alpha}\,\mathrm{d}H(\boldsymbol{\theta}) = \int_{\mathbb{S}_{+(\alpha)}^{d-1}}\sum_{i=1}^{d}\theta_i^{\alpha}\,\mathrm{d}H(\boldsymbol{\theta}) = \int_{\mathbb{S}_{+(\alpha)}^{d-1}}\mathrm{d}H(\boldsymbol{\theta}) = m.$$

In this thesis, all random vectors will be pre-processed to be on α-Fréchet margins and we take $b_n = n^{1/\alpha}$, so that

$$\begin{aligned}\sigma_{ii} &= \mathrm{scale}(X_i)^{\alpha} \\ &= \frac{\lim_{n\to\infty} n\mathbb{P}(X_i > n^{1/\alpha}x)}{x^{-\alpha}} \\ &= \frac{\lim_{n\to\infty} n\left\{1 - \exp\left[-(n^{1/\alpha}x)^{-\alpha}\right]\right\}}{x^{-\alpha}} \\ &= 1,\end{aligned}$$

and

$$m = \sum_{i=1}^{d} \sigma_{ii} = d.$$

Standardising the margins is akin to working with re-scaled variables with unit variance in the non-extremes setting. The appropriate analogue to the TPDM then becomes the correlation rather than covariance matrix.

As mentioned earlier, the TPDM's off-diagonal entries are simply pairwise EDMs. Thus the interpretation of σ_{ij} is inherited from the EDM: X_i and X_j are asymptotically independent if and only $\sigma_{ij} = 0$, and the magnitude of $\sigma_{ij} > 0$ reveals the strength of tail dependence between X_i and X_j. Like a correlation matrix, σ_{ij} attains its maximal value (one) when X_i and X_j are completely dependent (Example 2.7).

2.4.3 Decompositions of the TPDM

The TPDM is useful as a summary statistic for quantifying pairwise dependencies, but what sets it apart from other pairwise dependence matrices (e.g. the TDM)? The TPDM admits two types of decomposition: eigendecomposition and the completely positive decomposition (**cooley_decompositions_2019**). These underpin most statistical applications of the TPDM. The following results and proofs are reproduced from Kiriliouk and Zhou (2022).

Proposition 2.1. *The TPDM is symmetric and positive semi-definite.*

Proof. For any $i, j = 1, \ldots, d$,

$$\sigma_{ij} = \int_{\mathbb{S}_+^{d-1}} \theta_i^{\alpha/2} \theta_j^{\alpha/2} \, \mathrm{d}H(\boldsymbol{\theta}) = \int_{\mathbb{S}_+^{d-1}} \theta_j^{\alpha/2} \theta_i^{\alpha/2} \, \mathrm{d}H(\boldsymbol{\theta}) = \sigma_{ji}.$$

Hence $\Sigma = \Sigma^T$. For any $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$,

$$\boldsymbol{y}^T \Sigma \boldsymbol{y} = \boldsymbol{y}^T \mathbb{E}_H[\boldsymbol{\Theta}^{\alpha/2} (\boldsymbol{\Theta}^{\alpha/2})^T] \boldsymbol{y} = \mathbb{E}_H \left[\left(\boldsymbol{y}^T \boldsymbol{\Theta}^{\alpha/2} \right)^2 \right] \geq 0.$$

□

By standard linear algebra results, the TPDM can be decomposed as $\Sigma = UDU^T$, where $D \in \mathbb{R}^{d \times d}$ is a diagonal matrix of eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ and $U \in \mathbb{R}^{d \times d}$ is

an orthogonal matrix whose columns are the corresponding eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d \in \mathbb{R}^d$. *Foreshadow here.*

Definition 2.15. A matrix $M \in \mathbb{R}^{d\times d}$ is completely positive (CP) if there exists a matrix $B \in \mathbb{R}_+^{d\times q}$ such that $M = BB^T$.

Proposition 2.2. *The TPDM is completely positive.*

Proof. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ with angular measure H and TPDM Σ. By Proposition 5 in Fougères et al. (2013), there exists a sequence of matrices $\{A_q \in \mathbb{R}_+^{d\times q} : q \geq 1\}$ such that $H_q \xrightarrow{v} H$, where H_q is the angular measure of the max-linear random vector $\boldsymbol{X}_q \in \mathcal{RV}_+^d(\alpha)$ parametrised by A_q. The TPDM of $\boldsymbol{X}_q$ is $\Sigma_q = A_q^{\alpha/2}(A_q^{\alpha/2})^T$ by Example 2.7. Thus, $\{\Sigma_q : q \geq 1\}$ is a sequence of completely positive matrices. The limit $\lim_{q\to\infty} \Sigma_q = \Sigma$ must also be completely positive (CITE Theorem 2.2 in Berman & Shaked-Monderer (2003)).

□

In principle this provides a way to check whether a given matrix is a TPDM, but the membership problem for the completely positive cone is NP-hard (Dickinson and Gijben 2014). *Foreshadow here.*

2.4.4 The empirical TPDM

Definition 2.16. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ on Fréchet margins (2.31) and let H be the angular measure with respect to $\|\cdot\|_\alpha$ and normalising sequence $b_n = n^{1/\alpha}$. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be an iid sample of $\boldsymbol{X}$. The empirical TPDM estimator is the $d \times d$ matrix

$$\hat{\Sigma} = (\hat{\sigma}_{ij}), \qquad \hat{\sigma}_{ij} := \hat{E}_H[\Theta_i^{\alpha/2}\Theta_j^{\alpha/2}] = \frac{d}{k}\sum_{l=1}^{k} \Theta_{(l),i}^{\alpha/2}\Theta_{(l),j}^{\alpha/2}. \tag{2.56}$$

Note that the empirical TPDM implicitly depends on the customary tuning parameter k – or equivalently a radial threshold $t > 0$ – via the empirical angular measure.

Proposition 2.3. *The empirical TPDM is completely positive.*

Proof. Let $A = \hat{A}$, the $d \times k$ matrix with non-negative entries defined in (2.45). Then

$$\hat{A}^{\alpha/2}(\hat{A}^{\alpha/2})^T = \frac{d}{k}\sum_{i=1}^{k} \mathbf{\Theta}_{(i)}^{\alpha/2}\left(\mathbf{\Theta}_{(i)}^{\alpha/2}\right)^T = \hat{\Sigma}.$$

□

Proposition 2.4. *The empirical TPDM is symmetric and positive semi-definite.*

Proof. By complete positivity, $\hat{\Sigma} = AA^T$ for some matrix A. For any $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$,

$$\boldsymbol{y}^T\hat{\Sigma}\boldsymbol{y} = \boldsymbol{y}^T AA^T\boldsymbol{y} = \|A^T\boldsymbol{y}\|_2^2 \geq 0. \tag{2.57}$$

Since $\text{rank}(\hat{\Sigma}) = \text{rank}(AA^T) = \text{rank}(A)$, the empirical TPDM is positive definite if and only if the columns of A are linearly independent.

□

Proposition 2.5. *Under the conditions of Theorem 2.3, the entries of $\hat{\Sigma}$ are consistent and asymptotically normal, that is, for any $i, j = 1, \ldots, d$,*

$$\sqrt{k}(\hat{\sigma}_{ij} - \sigma_{ij}) \to \text{N}(0, \nu_{ij}^2), \qquad \nu_{ij}^2 := \text{Var}_H(\Theta_i^{\alpha/2}\Theta_j^{\alpha/2}). \tag{2.58}$$

Proof. See Example 2.4.

□

If X_i and X_j are asymptotically independent ($\sigma_{ij} = 0$), then $\nu_{ij}^2 = 0$ and the limit distribution is degenerate. In this case, the above result only proves consistency, i.e. $\hat{\sigma}_{ij} \to 0$. Thus, it is not possible to formally test for asymptotic independence ($\sigma_{ij} = 0$ against $\sigma_{ij} > 0$) using this result. Alternative strategies are explored in Lehtomaa and Resnick (2020)].

Using asymptotic normality one may construct asymptotic confidence intervals

$$\lim_{n\to\infty} \mathbb{P}\left[|\sigma_{ij} - \hat{\sigma}_{ij}| < z_{\beta/2}\sqrt{\nu_{ij}^2/k}\right] = 1 - \beta,$$

where $z_{\beta/2} = \Phi^{-1}(1-\beta/2)$. If the angular measure is known the asymptotic variance ν_{ij}^2 may be computed using the formula derived in Appendix XX.

Example 2.8. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ is symmetric logistic with $\gamma = 0.6$. Using Example 2.5 and results in Appendix XX, $\sigma_{ij} \approx 0.760$ and $\nu_{ij}^2 \approx 0.065$ for all $i \neq j$. For sufficiently large n,

$$\mathbb{P}\left[\hat{\sigma}_{ij} \in \left(0.760 \pm 1.96\sqrt{\frac{0.065}{k}}\right)\right] \approx 0.95.$$

For example, setting $n = 10^4$ and $k = \sqrt{n}$ yields $\mathbb{P}(0.710 < \hat{\sigma}_{ij} < 0.810) \approx 0.95$.

In practice, the asymptotic variance may be replaced with the plug-in estimator (Lee and Cooley 2023)

$$\hat{\nu}_{ij}^2 := \frac{1}{k-1}\sum_{l=1}^{k}\left(d\Theta_{(l),i}\Theta_{(l),j} - \hat{\sigma}_{ij}\right)^2.$$

The following result, proved by Krali (2018) for $\alpha = 2$, generalises asymptotic normality of the empirical TPDM to the entire matrix, rather than just individual entries. This is most simply expressed in terms of upper-half vectorisations of Σ and $\hat{\Sigma}$, that is

$$\begin{aligned}\boldsymbol{\sigma} &:= \text{vecu}(\Sigma) := (\sigma_{12}, \sigma_{13}, \ldots, \sigma_{1d}, \sigma_{23}, \ldots, \sigma_{2d}, \ldots, \sigma_{d-1,d}),\\ \hat{\boldsymbol{\sigma}} &:= \text{vecu}(\hat{\Sigma}) := (\hat{\sigma}_{12}, \hat{\sigma}_{13}, \ldots, \hat{\sigma}_{1d}, \hat{\sigma}_{23}, \ldots, \hat{\sigma}_{2d}, \ldots, \hat{\sigma}_{d-1,d}).\end{aligned}$$

Each vector contains $\binom{d}{2} = d(d-1)/2$ entries taken from the strictly upper-triangular part of each matrix. This is justified because the matrices are symmetric and we are not concerned with their diagonal entries. For simplicity, components are indexed according to the sub-indices of the corresponding matrix entry, e.g. the first entry of $\boldsymbol{\sigma}$ is σ_{12} rather than σ_1.

Proposition 2.6. *Under the conditions of Theorem 2.3, the estimator $\hat{\boldsymbol{\sigma}}$ is consistent and asymptotically normal, i.e.*

$$\sqrt{k}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) \to N(\mathbf{0}, V),$$

The diagonal and off-diagonal entries of the $\binom{d}{2} \times \binom{d}{2}$ asymptotic covariance matrix V are

given by

$$v_{ij,lm} := \lim_{k\to\infty} k\mathrm{Cov}(\hat{\sigma}_{ij}, \hat{\sigma}_{lm}) = \begin{cases} \nu_{ij}^2, & (i,j) = (l,m), \\ \rho_{ij,lm} & \textit{otherwise}, \end{cases}$$

where ν_{ij}^2 *is as defined in Proposition 2.5 and*

$$\rho_{ij,lm} := \frac{1}{2}\left[\mathrm{Var}_H(\Theta_i^{\alpha/2}\Theta_j^{\alpha/2} + \Theta_l^{\alpha/2}\Theta_m^{\alpha/2}) - \nu_{ij}^2 - \nu_{lm}^2\right].$$

The proof can be found in Appendix XX. It extends the proof of Theorem 5.23 in Krali (2018) to permit general α. The following example illustrates an application of Proposition 2.6 to the max-linear model.

Example 2.9. Suppose $\boldsymbol{X} = (X_1, \ldots, X_4) \in \mathcal{RV}_+^4(1)$ is max-linear with (randomly generated) parameter matrix $A \in \mathbb{R}_+^{4\times 12}$ as shown in Figure 2.3 (top). The TPDM $\Sigma = A^{1/2}(A^{1/2})^T$ is visualised in the bottom-left plot, with each cell's colour intensity representing the magnitude of the corresponding entry of Σ. All pairs of components exhibit strong dependence. The matrix in the bottom-right is the asymptotic covariance matrix V of $\hat{\boldsymbol{\sigma}}$, derived in Appendix XX. It has $\binom{4}{2} = 6$ rows and columns. *Any comments about the matrix itself?* We now run simulations verifying/illustrating Proposition 2.6 for this example. We generate $n = 10^4$ independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\boldsymbol{X} = A \times_{\max} \boldsymbol{Z}$ (see eq-max-linear-X) and compute the empirical TPDM using $k = \sqrt{n} = 100$ extremes. Repeating this process, we obtain 1,000 independent realisations of $\hat{\Sigma}$. After row-wise vectorisation, these estimates should be approximately $N(\boldsymbol{\sigma}, k^{-1}V)$ distributed. Figure 2.4 examines whether this is the case. First consider the diagonal panels. These show that the density function of an $N(\sigma_{ij}, \nu_{ij}^2/k)$ random variable (blue curve) provides a good fit for the empirical distribution of $\hat{\sigma}_{ij}$ (red histogram). Now consider the scatter plots in the lower triangular portion of the plot. The grey points represent 1,000 realisations of $(\hat{\sigma}_{ij}, \hat{\sigma}_{lm})$. The blue ellipses are the true asymptotic 95% data ellipses centred at $(\sigma_{ij}, \sigma_{lm})$ (blue crosses). Their orientation relates to the association $\rho_{ij,lm}$ between $\hat{\sigma}_{ij}$ and $\hat{\sigma}_{lm}$, while the lengths of the major and minor axes are dictated by the asymptotic variances ν_{ij}^2, ν_{lm}^2. The red ellipses and crosses are defined analogously but estimated from the data. They are generally in close agreement. The upper-triangular panels list the true values of $\rho_{ij,lm}$ (blue) alongside empirical estimates (red) based on the sample covariance between $\hat{\sigma}_{ij}$ and

$\hat{\sigma}_{lm}$.

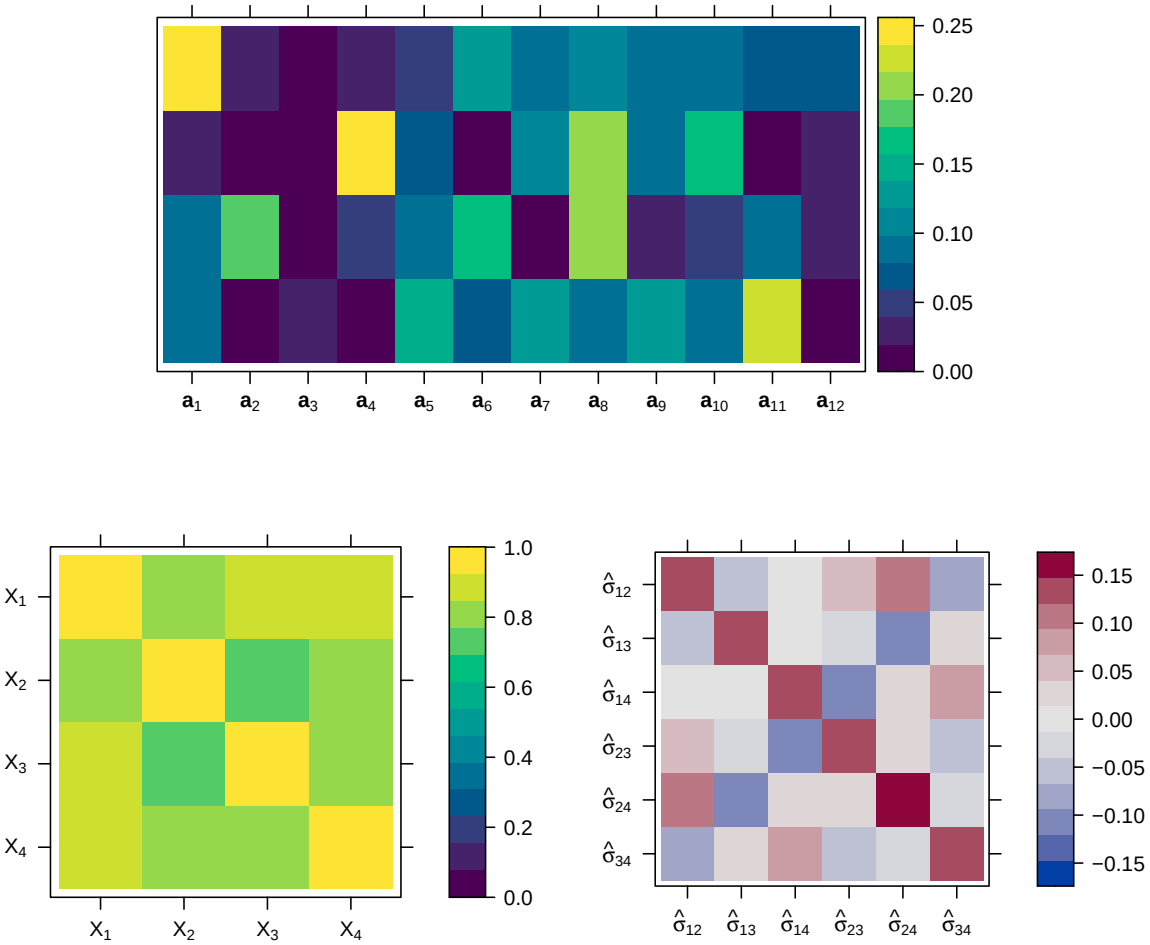

Figure 2.3: Visual representation of the matrices discussed in Example 2.9. Top: a randomly generated max-linear parameter matrix A with $d = 4$ and $q = 12$. Bottom left: the TPDM Σ of $\boldsymbol{X} = A \times_{\max} \boldsymbol{Z}$. Bottom right: the asymptotic covariance matrix V of $\hat{\boldsymbol{\sigma}}$.

2.5 Existing applications and extensions of the TPDM

The general goal of this thesis is to develop novel statistical applications of the TPDM for analysing extremal dependence. Before outlining our contributions, it seems logical to first familiarise the reader with existing TPDM-based methods in the literature. Our methods will either build upon these (e.g. compositional PCA in Chapter XXX) or address gaps in the field.

Our survey divides TPDM-related tools into four categories: principal components analysis

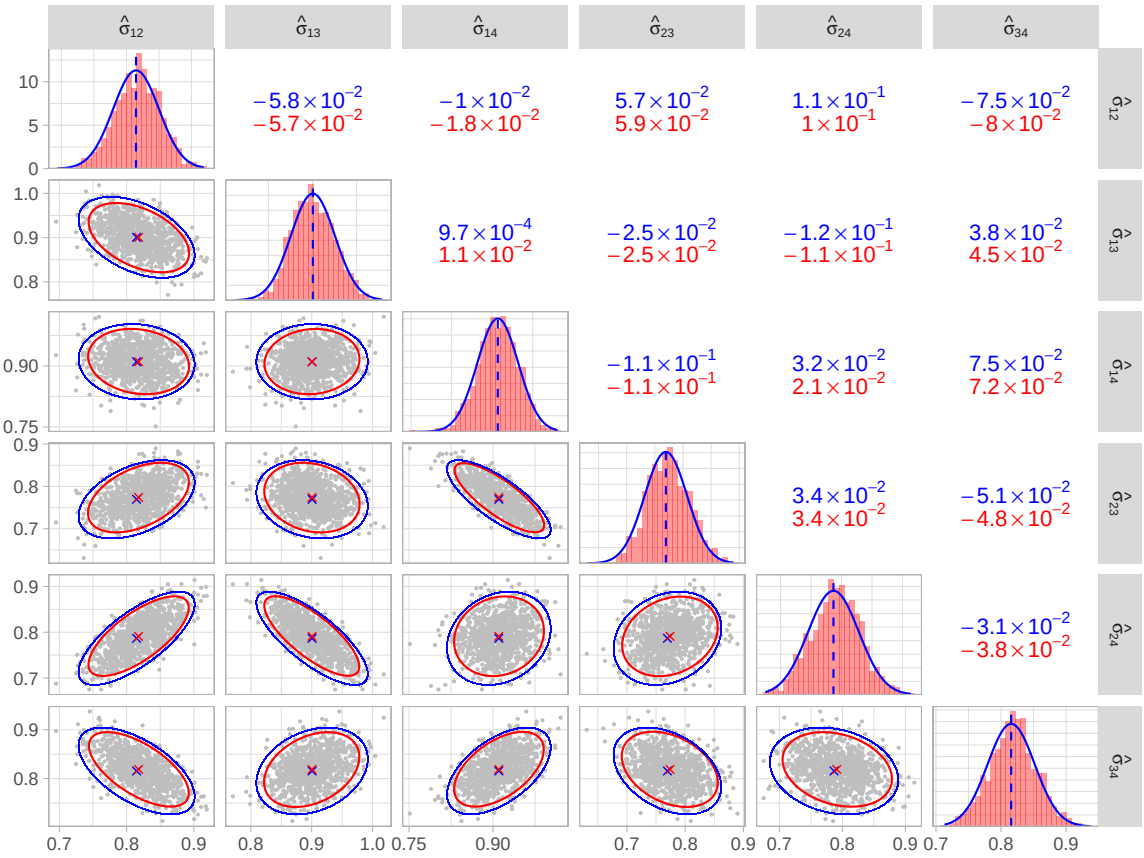

Figure 2.4: Pairs plot illustrating asymptotic normality of the empirical TPDM – see Example 2.9 for details. All panels: red represents the empirical quantity based on the 1,000 repeated simulations; blue represents the theoretical quantity based on asymptotic normality. Diagonal panels: the distribution (histogram or density function) of $\hat{\sigma}_{ij}$. Lower triangular panels: pairwise scatter plots of $(\hat{\sigma}_{ij}, \hat{\sigma}_{lm})$ (grey points) along with the mean (crosses) and the 95% data ellipse. Upper triangular panels: the entries $v_{ij,lm}$ of V.

(PCA), clustering, model fitting, and miscellaneous. The PCA methods leverage the TPDM eigendecomposition to perform dimension reduction. The extremal dependence structure is thereby represented by a low-dimensional object, facilitating exploratory analysis (Jiang et al. 2020; Russell and Hogan 2018; Szemkus and Friederichs 2024)

and generation of synthetic extreme events (Rohrbeck and Cooley 2023). The clustering techniques use the TPDM to partition a collection of random variables into groups according to asymptotic (in)dependence (Fomichov and Ivanovs 2023; Richards et al. 2024). When applied as a preliminary step, this splits a high-dimensional problem into several 'independent', low-dimensional, sub-problems. The model fitting section explains how the TPDM can be used to aid inference for the max-linear model (Fix et al. 2021; Kiriliouk

and Zhou 2022). Fitting a parametric model permits straightforward estimation of tail event probabilities. We conclude with a summary of other applications and extensions of the TPDM, such as in time series (Mhatre and Cooley 2021) and graphical models (Gong et al. 2024; Lee and Cooley 2023).

2.5.1 Principal component analysis (PCA) for extremes

Definition 2.17. The support of the angular measure has dimension $p^\star \ll d$.

This means the angular measure can be represented by a low-dimensional object, prompting the application of dimension reduction methods.

2.5.1.1 PCA in general finite-dimensional Hilbert spaces

In classical multivariate analysis, principal component analysis (PCA) is the flagship method for reducing the dimension of a random vector. PCA identifies linear subspaces that minimise the distance between the data and its low-dimensional projections.

PCA revolves around an underlying algebraic-geometric structure. Specifically, PCA assumes one is working in a Hilbert space $\mathcal{H}$. Without this theoretical foundation, it is meaningless to speak of principal components as orthogonal basis vectors or consider low-rank reconstructions as unique projections onto a subspace. A Hilbert space comprises a d-dimensional vector space with operations $\oplus$ and $\odot$ endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The induced norm and metric are $\| \cdot \|_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathcal{H}}^{1/2}$ and $d_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_{\mathcal{H}}$, respectively. In most applications $\mathcal{H} = \mathbb{R}^d$ with the usual Euclidean geometry. This thesis will additionally consider PCA in alternative spaces, including $\mathbb{R}^d_+$ and $\mathbb{S}^{d-1}_{+(1)}$. However, in each case, the Hilbert space in question will be isometric to the usual Euclidean space $(\mathbb{R}^d, \langle \cdot, \cdot \rangle)$. That is, there exists an isomorphism $h : \mathcal{H} \to \mathbb{R}^d$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{H}} = \langle h(\boldsymbol{x}), h(\boldsymbol{y}) \rangle, \qquad \|\boldsymbol{x} \ominus \boldsymbol{y}\|_{\mathcal{H}} = \|h(\boldsymbol{x}) - h(\boldsymbol{y})\|_2.$$

We present PCA for random vectors in $\mathbb{R}^d$, with the understanding that the data may have undergone an isometric transformation in pre-processing and outputs may need to be back-transformed to lie in the original space. This transform/back-transform approach is

$\mathcal{H}$	$\mathbb{R}^d$	$\mathbb{R}^d_+$ @cooleyDecompositionsDependenceHighdimensional2019	$\mathbb{S}^{d-1}_{+(1)}$ @aitchisonPrin
$h : \mathcal{H} \to \mathbb{R}^d$	$h(\boldsymbol{x}) = \boldsymbol{x}$	$h(\boldsymbol{x}) = \tau^{-1}(\boldsymbol{x}) = \log[\exp(\boldsymbol{x}) - 1]$	$h(\boldsymbol{x}) = \mathrm{clr}(\boldsymbol{x}) = \log[$
$h^{-1} : \mathbb{R}^d \to \mathcal{H}$	$h^{-1}(\boldsymbol{y}) = \boldsymbol{y}$	$h^{-1}(\boldsymbol{y}) = \tau(\boldsymbol{y}) = \log[1 + \exp(\boldsymbol{y})]$	$h^{-1}(\boldsymbol{y}) = \mathrm{clr}^{-1}(\boldsymbol{y}) =$
$\boldsymbol{x} \oplus \boldsymbol{y}$	$\boldsymbol{x} + \boldsymbol{y}$	$\tau[\tau^{-1}(\boldsymbol{x}) + \tau^{-1}(\boldsymbol{y})]$	$\mathcal{C}(x_1 y_1, \ldots, x_d y_d)$
$\alpha \odot \boldsymbol{x}$	$\alpha \boldsymbol{x}$	$\tau[\alpha \tau^{-1}(\boldsymbol{x})]$	$\mathcal{C}(x_1^\alpha, \ldots, x_d^\alpha)$
$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{H}}$	$\sum_{i=1}^d x_i y_i$	$\sum_{i=1}^d \tau^{-1}(x_i)\tau^{-1}(y_i)$	$\sum_{i=1}^d \log[x_i / \bar{g}(\boldsymbol{x})] \log$

equivalent to conducting the analysis in the original space with appropriately generalised notions of mean, variance, etc. (Pawlowsky-Glahn and Egozcue 2001).

Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ is a random vector in $\mathbb{R}^d$ satisfying $\mathbb{E}[\|\boldsymbol{Y}\|_2^2] < \infty$. Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be independent copies of $\boldsymbol{Y}$. The reconstruction error of a subspace $\mathcal{S} \subseteq \mathbb{R}^d$ is measured as

$$R(\mathcal{S}) := \mathbb{E}[\|\boldsymbol{Y} - \Pi_{\mathcal{S}} \boldsymbol{Y}\|_2^2] \tag{2.59}$$

Fundamental to PCA are the eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d \in \mathbb{R}^d$ and respective eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ of the positive semi-definite matrix

$$\Sigma = \mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^T].$$

The entries of Σ, herein referred to as the non-centred covariance matrix, are the second-order moments of $\boldsymbol{Y}$. By a change of basis, the random vector $\boldsymbol{Y}$ may be equivalently decomposed as

$$\boldsymbol{Y} = \sum_{j=1}^{d} \langle \boldsymbol{Y}, \boldsymbol{u}_j \rangle \, \boldsymbol{u}_j.$$

The scores $V_j := \langle \boldsymbol{Y}, \boldsymbol{u}_j \rangle$ represent the stochastic basis coefficients when $\boldsymbol{Y}$ is decomposed into the basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d\}$. They satisfy $\mathbb{E}[V_i V_j] = \lambda_i \mathbf{1}\{i = j\}$. For $1 \leq p < d$, the truncated expansion

$$\hat{\boldsymbol{Y}}^{[p]} := \sum_{j=1}^{p} V_j \boldsymbol{u}_j = \Pi_{\mathrm{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\}} \boldsymbol{Y}.$$

produces the optimal p-dimensional projection of $\boldsymbol{Y}$. In other words, the subspace $\mathcal{S}_p = \mathrm{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\}$ minimises the criterion (2.59) over $\mathcal{V}_p$, the set of all linear subspaces of dimension p of $\mathbb{R}^d$. It is the unique minimiser provided the multiplicity of λ_p is one. The corresponding risk is determined by the eigenvalues of the discarded components via $R(\mathcal{S}_p) = \sum_{j>p} \lambda_j$.

In practice, the covariance matrix is unknown so (2.59) cannot be minimised directly. Instead we resort to an empirical risk minimisation (ERM) approach, whereby the risk is replaced by

$$\hat{R}(\mathcal{S}) := \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{Y}_i - \Pi_{\mathcal{S}}\boldsymbol{Y}_i\|_2^2 \tag{2.60}$$

Minimisation of the empirical risk follows analogously based on the empirical non-centred covariance matrix

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y}_i \boldsymbol{Y}_i^T$$

and its ordered eigenpairs $(\hat{\lambda}_j, \hat{\boldsymbol{u}}_j)$ for $j = 1, \ldots, d$. For $p = 1, \ldots, d$ and $i = 1, \ldots, n$, the rank-p reconstruction of $\boldsymbol{Y}_i$ is given by

$$\hat{\boldsymbol{Y}}_i^{[p]} := \sum_{j=1}^{p} \hat{V}_{ij}\boldsymbol{u}_j = \Pi_{\text{span}\{\hat{\boldsymbol{u}}_1,\ldots,\hat{\boldsymbol{u}}_p\}}\boldsymbol{Y},$$

where $\hat{V}_{ij} := \langle \boldsymbol{Y}_i, \boldsymbol{u}_j \rangle$. The subspace $\hat{\mathcal{S}}_p = \text{span}\{\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_p\}$ minimises (2.60) in $\mathcal{V}_p$; the objective at the minimum is $\hat{R}(\hat{\mathcal{S}}_p) = \sum_{j>p} \hat{\lambda}_j$.

Usually the dimension of the target subspace (if it exists) is unknown, so the number of retained components p must be selected according to some criterion. At the heart of this choice is a trade-off between dimension reduction and approximation error. Selecting $p = \max\{j : \hat{\lambda}_j > 0\}$ results in perfect reconstructions but the reduction in dimension will be minimal if any. Excessive compression incurs information loss and destroys key features of the data. Several criteria for selecting the number of retained components based on the eigenvalues have been proposed. These include stopping when the reconstruction error $\sum_{j>p} \hat{\lambda}_j$ is acceptably small, cutting off components with $\lambda_j < 1$, or retaining components based on where the 'scree plot' forms an elbow.

If $\boldsymbol{Y}$ is mean-zero (or the $n \times d$ data matrix is column-centred in pre-processing), then Σ is the covariance matrix of $\boldsymbol{Y}$ and the procedure is termed centred PCA. In this case, PCA can be equivalently reformulated in terms of finding low-dimensional projections that maximally preserve variance. In the non-centred case this interpretation is not valid, the projections merely maximise variability around the origin. A detailed comparison between centred PCA and non-centred PCA is conducted in Cadima and Jolliffe (2009). They obtain relationships between and bounds on the eigenvectors/eigenvalues of the non-

centred and standard covariance matrices. Based on their theoretical analysis and a series of example, they conclude that both types of PCA generally produce similar results. In particular, the leading eigenvector (up to sign and scaling) of the non-centred covariance matrix is very often close to the vector of the column means of the data matrix. Thus the first non-centred principal component essentially relates to the centre of the data.

We now return to the context of multivariate extremes. Suppose $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ has sparse angular measure H and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is a sample of $\boldsymbol{X}$. There are several reasons why the the low-dimensional structure of the angular measure cannot be identified by naively applying standard PCA to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. At a practical level, the components $X_1, \ldots, X_d$ are heavy-tailed, so the requirement that second-order moments exist may be violated. The variance of an α-regularly varying random variable is infinite if $\alpha < 2$. More pertinently, standard PCA reveals relationships between variables in the centre rather than the tail of the joint distribution, because it arises from the covariance matrix. Moreover, the non-centred/centred covariance matrix captures dependence in both directions around the origin/mean, whereas we focus on extremes in a particular direction of interest ('positive'). Finally, standard PCA fails to capitalise on the probabilistic structure inherent to MRV random vectors. The one-dimensional radial component is (asymptotically) independent of the angular component. This points towards targetting dimension reduction at the angular component $\boldsymbol{\Theta}$ rather than the original vector $\boldsymbol{X}$. Indeed, the two key PCA methods of Drees and Sabourin (2021) and Cooley and Thibaud (2019) follow this approach. Despite emerging almost simultaneously, both are essentially based on eigendecomposition of the TPDM.

2.5.1.2 Drees and Sabourin (2021)

Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathrm{RV}_+^{d-1}(2)$ with angular measure H with respect to the Euclidean norm $\|\cdot\|_2$. The aim is to identify a low-dimensional linear subspace of $\mathbb{R}^d$ supporting H. For any subspace $\mathcal{S} \subset \mathbb{R}^d$, define the risk

$$R(\mathcal{S}) = \mathbb{E}_{\boldsymbol{\Theta} \sim H}[\|\boldsymbol{\Theta} - \Pi_{\mathcal{S}}\boldsymbol{\Theta}\|_2^2].$$

This represents the expected reconstruction error under the limit model. By assumption, there exists a linear subspace $\mathcal{S}^\star \in \mathcal{V}_{p^\star}$ of dimension $p^\star \ll d$ such that $R(\mathcal{S}^\star) = 0$, and $R(\mathcal{S}) > 0$ for all $\mathcal{S} \in \mathcal{V}_p$ with $p < p^\star$. The angular measure is unknown, so they adopt an ERM approach following the intuition that above a sufficiently high threshold the extremal angles will lie in a neighbourhood of $\mathcal{S}^\star$. The empirical risk is defined by replacing H with the empirical angular measure $\hat{H}$ based on the k largest observations in norm among a sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. That is

$$\hat{R}(\mathcal{S}) := \hat{\mathbb{E}}_{\boldsymbol{\Theta}\sim H}[\|\boldsymbol{\Theta} - \Pi_{\mathcal{S}}\boldsymbol{\Theta}\|_2^2] = \frac{m}{k}\sum_{i=1}^{k} \|\boldsymbol{\Theta}_{(i)} - \Pi_{\mathcal{S}}\boldsymbol{\Theta}_{(i)}\|_2^2.$$

This setup is almost identical to classical PCA on the random vector $\boldsymbol{\Theta}$. Note that boundedness of the simplex guarantees $\mathbb{E}[\|\boldsymbol{\Theta}\|_2^2] < \infty$. Let $\Sigma = \mathbb{E}_{\boldsymbol{\Theta}\sim H}[\boldsymbol{\Theta}\boldsymbol{\Theta}^T]$ be the TPDM of $\boldsymbol{X}$ and $(\boldsymbol{u}_j, \lambda_j)$ its (ordered) eigenpairs for $j = 1, \dots, d$. Then $\mathcal{S}_p = \operatorname{span}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_d\}$ minimises R in $\mathcal{V}_p$ and $R(\mathcal{S}_p) = \sum_{j>p} \lambda_p$. Choosing $p \geq p^\star$ yields $R(\mathcal{S}_p) = 0$. Analogously, the minimiser $\hat{\mathcal{S}}_p \in \mathcal{V}_p$ of $\hat{R}$ is the subspace spanned by the leading p eigenvectors of the empirical TPDM $\hat{\Sigma}$.

Drees and Sabourin (2021) derive theoretical statistical guarantees for their approach. Most importantly, they prove that the learnt subspace converges to the optimal one as the sample size increases to infinity. Provided $k(n)$ satisfies the rate conditions (2.43), then $\hat{\mathcal{S}}_p \to \mathcal{S}_p$ in the sense that

$$\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in\mathbb{S}_{+(2)}^{d-1}} \|\Pi_{\hat{\mathcal{S}}_p}\boldsymbol{\theta} - \Pi_{\mathcal{S}_p}\boldsymbol{\theta}\|_2 = 0.$$

If the target dimension is chosen correctly as $p = p^\star$, then $\hat{\mathcal{S}}_{p^\star} \to \mathcal{S}^\star$. They also provide high probability bounds on $|\hat{R}(\mathcal{S}) - R(\mathcal{S})|$ for fixed n.

Basing their approach on the angles viewed as points in $\mathbb{R}^d$ eases the derivation of theoretical guarantees, but creates interpretability issues. Consider $\hat{\boldsymbol{\Theta}}_i^{[p]} = \Pi_{\mathcal{S}_p}\boldsymbol{\Theta}_i$, the rank-$p$ reconstruction of an extremal angle $\boldsymbol{\Theta}_i$. Its components need not satisfy the unit-norm constraint and may even be negative. This may be remedied by shifting/normalising $\hat{\boldsymbol{\Theta}}_i^{[p]}$ appropriately, but its optimality properties will be destroyed in the process. One can also question whether Euclidean distances are an appropriate measure of angular reconstruction error; angular distances such as cosine distance may be better suited. Similarly, the hypothesis that the angular measure's low-dimensional structure manifests in a linear fash-

ion may be unrealistic, since data in the simplex are prone to exhibit curvature Aitchison (1983).

2.5.1.3 Cooley and Thibaud (2019)

The PCA technique developed by Cooley and Thibaud (2019) focusses on reconstruction and exploration of extreme events in terms of the original vector $\boldsymbol{X}$. As such, their PCA is grounded on an inner product space on $\mathcal{H} = \mathbb{R}^d_+$, the natural sample space of the data. The vector space is based on the softplus transformation

$$\tau : \mathbb{R} \to \mathbb{R}_+, \qquad \tau(x) = \log[1 + \exp(x)].$$

This transformation is bijective with inverse function $\tau^{-1}(y) = \log[\exp(y) - 1]$. The reason for choosing this particular mapping is that it is tail-preserving, i.e. $\lim_{x\to 1} \tau(x)/x = 1$. This provides an avenue for moving between the spaces $\mathbb{R}^d$ and $\mathbb{R}^d_+$ with negligible effect on the tails.

The linear-transformed inner product space is constructed as follows. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d_+$ and $\alpha \in \mathbb{R}$, define

$$\begin{aligned} \boldsymbol{x} \oplus \boldsymbol{y} &= \tau[\tau^{-1}(\boldsymbol{x}) + \tau^{-1}(\boldsymbol{y})] \\ \alpha \odot \boldsymbol{x} &= \tau[a\tau^{-1}(\boldsymbol{x})]. \end{aligned}$$

Then the vector space $(\mathbb{R}^d_+, \oplus, \odot)$ is endowed with an inner product and norm

$$\begin{aligned} \langle \boldsymbol{x}, \boldsymbol{y} \rangle_\tau &= \sum_{i=1}^{d} \tau^{-1}(x_i)\tau^{-1}(y_i) = \left\langle \tau^{-1}(\boldsymbol{x}), \tau^{-1}(\boldsymbol{y}) \right\rangle \\ \|\boldsymbol{x}\|_\tau &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle_\tau^{1/2} = \|\tau^{-1}(\boldsymbol{x})\|_2. \end{aligned}$$

The transform τ^{-1} is an isometry linking their inner product space on the positive orthant to the standard Euclidean space $\mathbb{R}^d$. Thus the PCA of Cooley and Thibaud (2019) can be equivalently formulated in $\mathbb{R}^d_+$ with regards to the original data in the space or in $\mathbb{R}^d$ using the transform/back-transform approach articulated earlier.

Suppose $\boldsymbol{X} \in \mathrm{RV}^d_+(\alpha)$ has TPDM Σ. Denote the ordered eigenpairs of Σ in $\mathbb{R}^d$ by $(\boldsymbol{u}_j, \lambda_j)$

for $j = 1, \ldots, d$. Then $\{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_d\} = \{\tau(\boldsymbol{u}_1), \ldots, \tau(\boldsymbol{u}_d)\}$ forms an orthonormal basis of $\mathbb{R}_+^d$. In this new basis, the random vector $\boldsymbol{X}$ may be decomposed as

$$\boldsymbol{X} = \bigoplus_{j=1}^{d} (V_j \odot \boldsymbol{\omega}_j) = \tau\left(\sum_{j=1}^{d} V_j \boldsymbol{u}_j\right),$$

where

$$V_j = \langle \boldsymbol{X}, \boldsymbol{\omega}_j \rangle_\tau = \left\langle \tau^{-1}(\boldsymbol{X}), \boldsymbol{u}_j \right\rangle, \qquad (j = 1, \ldots, d).$$

Rank-p reconstructions of $\boldsymbol{X}$ are obtained by the truncated expansion

$$\hat{\boldsymbol{X}}^{[p]} = \bigoplus_{j=1}^{d} (V_j \odot \boldsymbol{\omega}_j) = \tau\left(\sum_{j=1}^{d} V_j \boldsymbol{u}_j\right), \qquad (p = 1, \ldots, d).$$

The process follows analogously for PCA based on an independent sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and the empirical TPDM $\hat{\Sigma}$.

The elements of the $\mathbb{R}^d$-valued random vector $\boldsymbol{V} = (V_1, \ldots, V_d)$ are called the extremal principal components of $\boldsymbol{X}$. The random vector $\boldsymbol{V}$ is MRV with the same tail index as $\boldsymbol{X}$, but its angular measure H_V lives on the entire unit sphere, not just its restriction to the positive orthant. Although the dimension of $\boldsymbol{V}$ is the same as $\boldsymbol{X}$, the crucial difference is that its components are ordered according to their contribution to the extreme behaviour of $\boldsymbol{X}$. Proposition 6 in Cooley and Thibaud (2019) states that

$$\text{scale}(|V_i|) = \lambda_i^{1/\alpha}, \qquad (i = 1, \ldots, d),$$

and therefore $\text{scale}(|V_1|) \geq \ldots \geq \text{scale}(|V_d|) \geq 0$. The ith eigenvector $\boldsymbol{\omega}_i$ represents the direction of maximum scale after accounting for information contained in $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{i-1}$; sequential examination of the eigenvectors provides insight into the extremal dependence structure.

2.5.1.4 Applications

The PCA method of Cooley and Thibaud (2019) has been applied for exploratory purposes in the context of climatology (Jiang et al. 2020; Szemkus and Friederichs 2024), finance (Cooley and Thibaud 2019) and sport (Russell and Hogan 2018).

Jiang et al. (2020) analyse the extremal behaviour of precipitation across the United States. They discover an increasing temporal trend in the coefficient of the first principal component V_1, and relate the eigenvectors to the El-Niño Southern Oscillation (ENSO), a cyclical phenomenon that is known to be a key climatological driver. They find that low-rank reconstructions of Hurricane Floyd broadly capture the event's large-scale structure, but a large number of eigenvectors are needed to recreate more localised features. The spatial extent of the study region and relatively localised behaviour of extreme behaviour leads them to consider a 'pairwise-thresholded' estimator of the TPDM instead of the usual estimator (2.56) thresholded on the norm of entire vector. This alternative estimator is given by

$$\tilde{\Sigma} = (\tilde{\sigma}_{ij}), \qquad \tilde{\sigma}_{ij} = \frac{2}{k} \sum_{l=1}^{n} \Theta_{li}\Theta_{lj} \mathbf{1}\{R_l^{ij} > R_{(k+1)}^{ij}\},$$

where $R_l^{ij} = \|(X_{li}, X_{lj})\|$ and $R_{(k+1)}^{ij}$ is the $(k+1)$th upper order statistic of $\{R_l^{ij} : l = 1, \ldots, n\}$. The estimator $\tilde{\Sigma}$ is not positive semi-definite, so the PCA analysis is instead conducted using the nearest positive definite matrix in Frobenius norm. The ramifications of this ad-hoc step, in terms of the estimator's theoretical properties and practical performance, are not studied.

Szemkus and Friederichs (2024) devise an extension of the TPDM, called the cross-TPDM, to study the joint extremal behaviour between two sets of variables. They analyse two meteorological variables – daily maximum temperature and a measure of accumulated precipitation deficit – to describe the dynamics of summer heatwaves in Europe. The cross-TPDM is the analogue of the cross-covariance matrix. Letting $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathrm{RV}_+^p(2)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_q) \in \mathrm{RV}_+^q(2)$, the cross-TPDM is defined as the $p \times q$ matrix with entries

$$\sigma_{ij}^{XY} = \int_{\mathbb{S}_+^{p+q-1}} \theta_i^X \theta_j^Y \, \mathrm{d}H(\boldsymbol{\theta}),$$

where H is the angular measure of $(\boldsymbol{X}, \boldsymbol{Y}) = (X_1, \ldots, X_p, Y_1, \ldots, Y_q) \in \mathrm{RV}_+^{p+q}(2)$ and the variable of integration is indexed as $\boldsymbol{\theta} = (\theta_1^X, \ldots, \theta_p^X, \theta_1^Y, \ldots, \theta_q^Y)$. (This definition could be extended to cater for an arbitrary tail index by introducing the usual $\alpha/2$ exponents in the integrand.) In the context of their climatological study, the entry σ_{ij}^{XY} represents the strength of extremal dependence between the maximum temperature at location i and the precipitation deficit at location j. The singular-value decomposition of the cross-TPDM is

used to analyse the dynamics of compound extreme events. They devise extremal pattern indices to quantify whether particular patterns of interest – those signified by the singular vectors of the cross-TPDM – are highly pronounced.

A more unusual application of the TPDM is found in Russell and Hogan (2018). Their study characterises the difference in performance between typical and elite-level National Football League (NFL) performers across the Scouting Combine event. The Combine comprises six physical tests: Bench Press, Vertical Jump, Broad Jump, 40-yard Sprint, the Shuttle Drill, and the Three Cone Drill. The tests afford teams the opportunity to gauge the athletic ability of prospective players, thereby influencing whether (or how highly) they are drafted for the upcoming season. Russell and Hogan (2018) explore how strongly player performance correlates across these tests. Intuitively, if two events exhibit strong association, then they may be measuring the same underlying skills (speed, strength, agility etc.). After standardising player performance to account for differences in playing position, they find significant differences between the bulk dependence structure and the extremal dependence structure. In particular, the leading eigenvectors of the covariance matrix reveal that the Combine events cluster into three distinct groups, corresponding to strength, agility, and explosiveness. On the other hand, the TPDM eigenvectors produce only two such groups: power and agility. This reveals differences between non-elite and elite performers; recommendations regarding the composition of the Combine events are made accordingly.

Rohrbeck and Cooley (2023) move beyond the use of the extremal PCA for purely exploratory purposes and demonstrate how it be used to generate synthetic extreme events. Hazard event sets are widely used in catastrophe modelling to assess exposure to extreme events. Imagine an insurance company insures against damage to a portfolio of properties, and wishes to gauge its exposure to claims caused by flooding. Given (i) the spatial locations of these properties, (ii) other relevant characteristics such as property value and construction standard, and (iii) a set of simulated flood events, one can derive a probabilistic loss distribution. If the exposure is unacceptably high, they might adjust their underwriting strategy or purchase reinsurance. Rohrbeck and Cooley (2023) show how to generate approximate samples from H, even in high-dimensions, by leveraging the PCA method of Cooley and Thibaud (2019). Their generative framework hinges on the fact that

the leading components of $\boldsymbol{V}$ account for the greatest proportion of extremal behaviour of $\boldsymbol{X}$. Thus, efforts may be concentrated towards modelling the dependence structure of the sub-vector $(V_1, \ldots, V_p)$ for some appropriately chosen $p < d$. To achieve this, they use a spherical kernel density estimate to flexibly model the dependence between $V_1, \ldots, V_p$ and additionally between $(V_1, \ldots, V_p)$ and $(V_{p+1}, \ldots, V_d)$. The dependence structure of $(V_{p+1}, \ldots, V_d)$ is simply modelled by a nearest-neighbours approach. The number of components p entering into the complex model is selected by a leave-one-out cross validation procedure. This involves discarding an extreme observation $\boldsymbol{x}_{(i)}$, generating a large number of samples $\tilde{\boldsymbol{x}}_1^{[p]}, \ldots, \tilde{\boldsymbol{x}}_N^{[p]}$ for a range of values p, and then assessing whether any of the generated samples resemble the discarded event using

$$D_i(p) = \min_{l=1,\ldots,N} \varrho\left(\boldsymbol{x}_{(i)}, \tilde{\boldsymbol{x}}_l^{[p]}\right),$$

where $\varrho(\cdot, \cdot)$ is an angular dissimilarity measure. After repeating for all extreme events $i = 1 \ldots, k$, one chooses the optimal p as that which minimises the average error

$$\bar{D}(p) = \frac{1}{k} \sum_{i=1}^{k} D_i(p).$$

Their approach is illustrated using historical river flow data across $d = 45$ gauges in northern England and southern Scotland. They select $p = 7$ and find reasonable agreement between the observed river flow extreme events and the synthetic ones generated by their algorithm, e.g. by examining QQ-plots comparing the observed and sampled distributions of $\max_{j \in \mathcal{G}} X_j$ or $\|(X_i : i \in \mathcal{G})\|$ for selected groups of gauges $\mathcal{G} \subset \{1, \ldots, d\}$.

Add more critical comments, especially about asymptotic independence when using large study regions or with localised extremes, e.g. rainfall. Or leave this to the 'bias' section?

2.5.2 Clustering into asymptotically dependent groups

Within multivariate extremes, the umbrella term `clustering` can refer to a multitude of tasks. To avoid confusion, we briefly describe these and clarify which type we are referring to.

- **Prototypical events.** Assume that the angular measure concentrates at/near a

small number of points in $\mathbb{S}_+^{d-1}$. Then one might wish to identify cluster centres $\boldsymbol{w}_1, \ldots \boldsymbol{w}_K$ minimising some objective function of the form

$$\mathbb{E}_{\boldsymbol{\Theta} \sim H} \left[\min_{l=1,\ldots,K} \varrho(\boldsymbol{\Theta}, \boldsymbol{w}_l) \right], \tag{2.61}$$

where $\varrho : \mathbb{S}_+^{d-1} \times \mathbb{S}_+^{d-1} \to [0,1]$ is some distance/dissimilarity function. The cluster centres can be interpreted as the directions of prototypical extremes events. See Chautru (2015), Janßen and Wan (2020) and Medina et al. (2021) for further details.

- **Identification of concomitant extremes.** Suppose that angular measure is supported on a set of $K \ll 2^{d-1}$ subspaces (faces) of the simplex $C_{\beta_1}, \ldots, C_{\beta_K}$, where $\beta_1, \ldots, \beta_K \in \mathcal{P}(\{1, \ldots, d\}) \setminus \emptyset$ and

$$C_\beta = \{\boldsymbol{\theta} \in \mathbb{S}_+^{d-1} : \theta_i > 0 \iff i \in \beta\}.$$

  Only those groups ('clusters') of components indexed by $\beta_1, \ldots, \beta_K$ may be simultaneously extreme. Identification of the support of the angular measure is notoriously challenging because the extremal angles $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}$ lie (almost surely) in the interior of the simplex. Goix et al. (2017) and Simpson et al. (2020) identify clusters according to whether observations fall within appropriately sized rectangular/conic neighbourhoods of the corresponding axis in $\mathbb{R}_+^d$. Meyer and Wintenberger (2020) take a different approach, whereby the angular component is defined with respect to the Euclidean projection (Liu and Ye 2009) rather than usual projection based on self-normalisation. The geometry of the projection is such that the projected data lie on subfaces of the simplex. The price paid is that the limiting conditional distribution of the angles is related to, but not identical to, the angular measure.
- **Partitioning into AD/AI groups components.** This notion of clustering is related to the previous type. We assume that the variables $X_1, \ldots, X_d$ can be partitioned into K clusters, such that X_i and X_j are asymptotically dependent if and only if they belong to the same cluster. In other words, there exists $2 \leq K \leq d$ and a partition $\beta_1, \ldots, \beta_K$ of $\{1, \ldots, d\}$ such that the angular measure is supported on

$C_{\beta_1}, \ldots, C_{\beta_K}$ or lower-dimensional subspaces thereof, i.e.

$$H\left(\bigcup_{l=1}^{K} \bigcup_{\beta_l' \subseteq \beta_l} C_{\beta_l'}\right) = m.$$

The task of modelling the dependence structure of $\boldsymbol{X}$ can be divided into lower-dimensional sub-problems involving the random sub-vectors $\boldsymbol{X}_{\beta_1}, \ldots, \boldsymbol{X}_{\beta_K}$. If $K = d$, then all variables are asymptotically independent. The underlying hypothesis is very strong and unlikely to hold in practice. Nevertheless, it is often a useful simplifying modelling assumption. Bernard et al. (2013) propose grouping components using the k-medoids algorithm (Kaufman and Rousseeuw 1990) with a dissimilarity matrix populated with pairwise measures of tail dependence, similar to χ_{ij} and σ_{ij}. The approaches of Fomichov and Ivanovs (2023) and Richards et al. (2024) involve the TPDM; these are reviewed in greater detail below.

2.5.2.1 Fomichov and Ivanovs (2023)

Fomichov and Ivanovs (2023) show that the latter kind of clustering may be performed using the framework of the first kind. They provide a link between the principal eigenvector $\boldsymbol{u}_1$ of the TPDM and the minimiser of the objective (2.61) with quadratic cost $\varrho(\boldsymbol{\theta}, \boldsymbol{\phi}) = \langle \boldsymbol{\theta}, \boldsymbol{\phi} \rangle^2$ and $K = 1$:

$$\min_{\boldsymbol{\theta} \in \mathbb{S}_{+(2)}^{d-1}} \mathbb{E}_{\boldsymbol{\Theta} \sim H} \left[\varrho(\boldsymbol{\Theta}, \boldsymbol{\theta})\right] = \mathbb{E}_{\boldsymbol{\Theta} \sim H} \left[\varrho(\boldsymbol{\Theta}, \boldsymbol{u}_1)\right].$$

Note that $\boldsymbol{u}_1 \in \mathbb{S}_{+(2)}^{d-1}$ is assumed to be suitably normalised with all entries being non-negative; the Perron-Frobenius theorem guarantees this is possible. This result informs an iterative clustering procedure called spherical k-principal-components. Consider a set of extremal angles $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(k)} \in \mathbb{S}_{+(2)}^{d-1}$ and current centroids $\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_K \in \mathbb{S}_{+(2)}^{d-1}$. A single iteration of their procedure yields new centroids $\hat{\boldsymbol{w}}_1^\star, \ldots, \hat{\boldsymbol{w}}_K^\star \in \mathbb{S}_{+(2)}^{d-1}$ given by the respective principal eigenvectors of

$$\hat{\Sigma}^{[i]} = \sum_{l=1}^{k} \boldsymbol{\theta}_{(l)} \boldsymbol{\theta}_{(l)}^T \mathbf{1}\{\arg\min_{j=1,\ldots,K} \varrho(\boldsymbol{\theta}_{(l)}, \boldsymbol{w}_j) = i\}, \qquad (i = 1, \ldots, K).$$

The matrix $\hat{\Sigma}^{[i]}$ represents the empirical TPDM (up to some multiplicative constant) based on the nearest neighbours of the ith centroid. Fomichov and Ivanovs (2023) prove that, under certain conditions, the limiting centroids lie in a neighbourhood of the faces of interest $C_{\beta_1}, \ldots, C_{\beta_K}$. Thresholding the centroid vectors yields the final partition $\beta_1, \ldots, \beta_K$.

2.5.2.2 Richards et al. (2024)

Richards et al. (2024) apply hierarchical clustering using the empirical TPDM as the underlying similarity matrix. The clustering method constitutes a minor aspect of their submission to the EVA (2023) Data Challenge. Few methodological details are provided, so the following explanation constitutes our interpretation of their method, drawing on Figure 4 in Richards et al. (2024) and the accompanying code made available at https://github.com/matheusguerrero/yalla. Define the dissimilarity between X_i and X_j as $\varrho_{ij} = 1-\sigma_{ij}$. This satisfies the properties of a dissimilarity measure (CITE: A MATHEMATICAL THEORY FOR CLUSTERING IN METRIC SPACES):

$$\varrho_{ij} \geq 0, \qquad \varrho_{ii} = 0, \qquad \varrho_{ij} = \varrho_{ji}.$$

The $d \times d$ dissimilarity matrix $\mathcal{D} = 1 - \Sigma = (\varrho_{ij})$ can be fed into standard hierarchical clustering algorithms. Agglomerative hierarchical clustering initially assigns each variable belongs to its own cluster, i.e. $\beta_i = \{i\}$ for $i = 1, \ldots, d$. The algorithm proceeds iteratively, repeatedly joining together the two closest clusters until some stopping criterion is satisfied. Under complete-linkage clustering, the distance between clusters $\beta \neq \beta'$ is given by $\max\{\varrho_{ij} : i \in \beta, j \in \beta'\}$. The merging process may be stopped when there is a sufficiently small number of clusters or when the clusters are sufficiently separated.

2.5.3 Parametric model fitting

Inference for max-linear models (i.e. estimating the parameter matrix A) is a challenging task. The lack of an angular density function precludes the use of standard maximum likelihood procedures. Einmahl, Kiriliouk, et al. (2018) propose a procedure that minimises a weighted least-squares distance to some initial (non-parametric) estimator. Their procedure becomes computationally intensive when q is large. Janßen and Wan (2020) and

Medina et al. (2021) cluster the angles of extreme observations and identify the normalised columns of A with the q cluster centres. The minimum-distance and clustering approaches assume q is fixed; Kiriliouk (2020) present a hypothesis test to assist with choosing q. *Alternative strategies based on TPDM etc. are explained later; do I move this section to there?*

Fix et al. (2021) consider the extremal behaviour of a spatial process $\{\boldsymbol{X}(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^2\}$ at fixed sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d \in \mathbb{R}^2$ by modelling $\boldsymbol{X} = (\boldsymbol{X}(\boldsymbol{s}_i) : i = 1 \ldots, d) \in \mathrm{RV}_+^d(2)$ as

$$\boldsymbol{X} = (I - \rho W)^{-1} \otimes \boldsymbol{Z}. \tag{2.62}$$

This is called the extremal spatial auto-regressive (SAR) model. The $d \times d$ matrix W contains the (known) pairwise spatial distances and $\rho \in (0, 1/4)$ is a spatial dependence parameter. The extremal SAR model is a special case of the max-linear model (2.22) with $A = A(\rho) = (I - \rho W)^{-1}$. They propose estimating the model parameter ρ by minimising the discrepancy between the empirical TPDM and the theoretical TPDM $\Sigma(\rho) := A(\rho)A(\rho)^T$, that is

$$\hat{\rho} = \underset{\rho \in (0,1/4)}{\arg\min} \|\hat{\Sigma} - \Sigma(\rho)\|_F^2. \tag{2.63}$$

In fact, $\hat{\Sigma}$ is replaced with a bias-corrected version of the empirical TPDM; this will be discussed in Section XX.

Kiriliouk and Zhou (2022) consider the more general problem of modelling arbitrary max-linear random vectors $\boldsymbol{X} = A \times_{\max} \boldsymbol{Z} \in \mathrm{RV}_+^d(2)$. In a similar spirit to Fix et al. (2021), they propose estimating A so as to enforce conformity between the empirical and model TPDMs. This means that the estimate of A belongs to the set

$$\mathcal{CP}(\hat{\Sigma}) := \left\{ \hat{A} \in \mathbb{R}_+^{d \times q} : q \geq 1,\ \hat{\Sigma} = \hat{A}^{\alpha/2}(\hat{A}^{\alpha/2})^T \right\}.$$

Choosing $\hat{A} \in \mathcal{CP}(\hat{\Sigma})$ guarantees that the pairwise dependencies of the fitted model match those exhibited by the data. The set $\mathcal{CP}(\hat{\Sigma})$ is in direct correspondence to the set of completely positive (CP) factors of $\hat{\Sigma}$; we call $\hat{A} \in \mathcal{CP}(\hat{\Sigma})$ a CP-estimate of A. The naive estimate (2.45) belongs to this class, but Kiriliouk and Zhou (2022) provide an algorithm for efficiently obtaining further estimates $\hat{A} \in \mathbb{R}_+^{d \times d} \cap \mathcal{CP}(\hat{\Sigma})$.

Fix et al. (2021) and Kiriliouk and Zhou (2022) evaluate the practical performance of their estimators by computing tail event probabilities in a series of simulated/real-world scenarios. *More details here, when I've written up formulae for failure events.*

2.5.4 Miscellaneous: time series and extremal graphical models

(Mhatre and Cooley 2021; Gong et al. 2024; Lee and Cooley 2023).

2.6 Bias in the empirical TPDM in weak-dependence scenarios}

Section XX reviewed the asymptotic properties of the empirical TPDM. We recall in particular that it is asymptotically unbiased, meaning $\mathbb{E}[\hat{\Sigma}] \to \Sigma$ as $n \to \infty$. The associated rate of convergence is $\mathcal{O}(k^{-1/2})$, where k represents the number of extreme observations and satisfies the rate conditions (2.43). For example, choosing $k(n) = \sqrt{n}$ yields a convergence rate of $\mathcal{O}(n^{-1/4})$. In practical settings the number of extreme events k is normally small, both in relative (by definition) and absolute terms. For example, commonly available climate records typically span approximately 50 years (**boulaguiem_modeling_2022**). A study of temperature extremes might then be based on, say, $n \approx 50 \times 100 = 5,000$ daily observations recorded in the summer months over this time span. Working with small effective sample sizes means it is critical to understand the non-asymptotic, finite-sample performance of the empirical TPDM.

2.6.1 Bias in threshold-based estimators

At finite levels, the empirical TPDM exhibits an upwards bias in weak dependence scenarios (Cooley and Thibaud 2019; Fix et al. 2021; Mhatre and Cooley 2021). This is true more generally of threshold-based estimators in multivariate extremes (Raphaël Huser et al. 2016). They conduct simulation studies with $d = 2$ and $n = 10^4$ examining the performance of various estimators of γ, the dependence parameter of the symmetric logistic model. The results show that block-maxima based estimators have a small bias but very high variability. On the other hand, each of the threshold-based estimators $\hat{\gamma}$ tend to overestimate the

dependence strength, that is $\text{Bias}(\hat{\gamma}) = \mathbb{E}[\hat{\gamma}] - \gamma < 0$. Moreover, the discrepancy increases as dependence weakens ($\gamma \to 1$).

The empirical TPDM suffers from the same issue when dependence is weak. This can be summarised as

$$\sigma_{ij} \ll 1 \implies \text{Bias}(\hat{\sigma}_{ij}) = \mathbb{E}[\hat{\sigma}_{ij}] - \sigma_{ij} > 0. \tag{2.64}$$

Note that overestimating the dependence strength now corresponds to a positive bias, so the inequality is reversed.

2.6.2 Simulation experiments

See Figure 2.5.

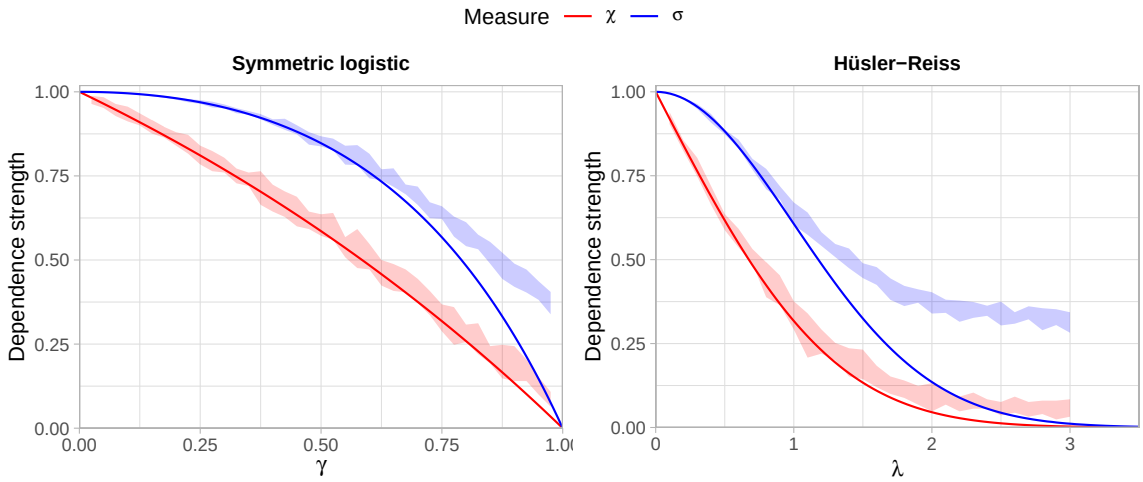

Figure 2.5: True dependence strengths for the symmetric logistic (left) and Hüsler-Reiss (right) models, measured using the tail dependence coefficient (red line) and TPDM (blue line). The shaded regions represent the minimum/maximum values of empirical estimates over 10 repeated simulations using bivariate samples of size $n = 5 \times 10^3$.

2.6.3 Existing approaches to bias-correction for the TPDM

Estimation error in the empirical TPDM was first studied by Cooley and Thibaud (2019). In the Supplementary Material, they assess the accuracy of the eigenvalues/eigenvectors of the empirical TPDM. Their example is based on a Brown-Resnick process, for which the true TPDM is known (Example 2.6). They find that the leading eigenvalue is overestimated

($\hat{\lambda}_1 > \lambda_1$) and subsequent eigenvalues are underestimated ($\hat{\lambda}_j < \lambda_j$ for $j \geq 2$). The bias reduces when the sample size and radial threshold are increased. In the non-extreme setting, the sample covariance matrix has the same deficiency, especially when the sample size and dimension are comparable in magnitude (Mestre 2008). Poor spectrum estimation can have important consequences in a downstream analysis, such as deciding how many principal components are retained in PCA. Cooley and Thibaud (2019) do not propose any solutions to improve TPDM estimation.

The bias issue (2.64) is addressed more directly by Mhatre and Cooley (2021). In their Supplementary Material, they conduct simulation studies to examine the performance of the empirical TPDF, the time series analogue of the TPDM (see Section XX). They show that $\sigma(h)$ exhibits a positive bias, especially at higher lags where the theoretical TPDM should vanish to zero. Their bias-corrected TPDF estimator works by subtracting the mean from the time series in pre-processing. The rationale for their estimator is described in terms of the position of extreme points in a lag plot (i.e. a scatter plot of (X_t, X_{t+h}) for some fixed lag h). Subtracting the mean has little effect on points near the middle of this plot, but points close to the coordinate axes are driven even closer.

The first bias-correction estimation procedure for the TPDM is found in Fix et al. (2021). Recall from Section XX that they use the empirical TPDM to estimate the spatial dependence parameter ρ of the extremal SAR model (2.62). When the spatial extent of the study domain is large compared to that of the modelled phenomenon, their estimation procedure (2.63) is liable to overestimate ρ. This is because the empirical TPDM fails to capture the weak dependence between distant pairs of sites. Their bias-correction procedure is founded on the assumption that the pairwise asymptotic dependence strength vanishes to zero as the distance between two sites increases. Consider a spatial process $\{X(\boldsymbol{s}) : s \in \mathbb{R}^2\}$ and fixed locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d \in \mathbb{R}^2$. Let $X_i = X(\boldsymbol{s}_i)$ represent the process at site i and h_{ij} the spatial distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. Treating the empirical TPDM entries as functions of distance, they model the relationship between the empirical TPDM and spatial distance via

$$\hat{\sigma}(h) = \beta_0 \exp(-\beta_1 h) + \beta_2.$$

The parameters $\beta_0, \beta_1, \beta_2$ are estimated from the observed data $\{(\hat{\sigma}_{ij}, h_{ij}) : 1 \leq i < j \leq d\}$ by non-linear least squares estimation, e.g. using `nls()`. Since $\hat{\sigma}(h) \to \beta_2$ as $h \to \infty$, the

horizontal asymptote $\hat{\beta}_2$ of the fitted model is used as a proxy for the bias at large distances. This determines the amount of shrinkage that should be applied to the off-diagonal entries, yielding the final estimator

$$\tilde{\Sigma} = (\tilde{\sigma}_{ij}), \qquad \tilde{\sigma}_{ij} = \begin{cases} \hat{\sigma}_{ij}, & i = j, \\ (\hat{\sigma}_{ij} - \hat{\beta}_2)_+, & i \neq j. \end{cases} \tag{2.65}$$

Estimates of the diagonal entries are found to be unbiased – and their values are known if the margins are standardised – so they are unaltered. Fix et al. (2021) find that $\tilde{\Sigma}$ is effective in reducing the bias in estimation of ρ. Its performance more broadly as an estimator for Σ is not studied. In any case, their procedure is only applicable in settings where there is a notion of distance between variables. The estimator (2.65) results from element-wise application of the soft-thresholding operator (with shrinkage parameter $\hat{\beta}_2$) to the empirical TPDM (**rothman_generalized_2009**). This connection will be developed further in Chapter XX, where we propose alternative bias-corrected TPDM estimators.

References

Aitchison, J. (1983). "Principal Component Analysis of Compositional Data". In: *Biometrika* 70.1, pp. 57–65.

Bernard, Elsa et al. (2013). "Clustering of Maxima: Spatial Dependencies among Heavy Rainfall in France". In: *Journal of Climate* 26.20, pp. 7929–7937.

Brown, B. M. and Sidney Resnick (1977). "Extreme Values of Independent Stochastic Processes". In: *Journal of Applied Probability* 14.4, pp. 732–739.

Cadima, Jorge and Ian Jolliffe (2009). "On Relationships Between Uncentred and Column-Centred Principal Component Analysis". In: *Pakistan Journal of Statistics* 25.4, pp. 473–503.

Chautru, Emilie (2015). "Dimension Reduction in Multivariate Extreme Value Analysis". In: *Electronic Journal of Statistics* 9.1, pp. 383–418.

Clémençon, Stéphan et al. (2023). "Concentration Bounds for the Empirical Angular Measure with Statistical Learning Applications". In: *Bernoulli* 29.4.

Coles, Stuart, Janet Heffernan, and Jonathan Tawn (1999). "Dependence Measures for Extreme Value Analyses". In: *Extremes* 2.4, pp. 339–365.

Coles, Stuart and J A Tawn (1994). "Statistical Methods for Multivariate Extremes: An Application to Structural Design". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43.1, pp. 1–48.

Cooley, Daniel and Emeric Thibaud (2019). "Decompositions of Dependence for High-Dimensional Extremes". In: *Biometrika* 106.3, pp. 587–604.

Davison, A. C., S. A. Padoan, and M. Ribatet (2012). "Statistical Modeling of Spatial Extremes". In: *Statistical Science* 27.2, pp. 161–186.

Dickinson, Peter J. C. and Luuk Gijben (2014). "On the Computational Complexity of Membership Problems for the Completely Positive Cone and Its Dual". In: *Computational Optimization and Applications* 57.2, pp. 403–415.

Dombry, Clément, Sebastian Engelke, and Marco Oesting (2016). "Exact Simulation of Max-Stable Processes". In: *Biometrika* 103.2, pp. 303–317.

Drees, Holger and Anne Sabourin (2021). "Principal Component Analysis for Multivariate Extremes". In: *Electronic Journal of Statistics* 15.1, pp. 908–943.

Einmahl, John H. J., Anna Kiriliouk, and Johan Segers (2018). "A Continuous Updating Weighted Least Squares Estimator of Tail Dependence in High Dimensions". In: *Extremes* 21.2, pp. 205–233.

Einmahl, John H. J., Andrea Krajina, and Johan Segers (2012). "An M-estimator for Tail Dependence in Arbitrary Dimensions". In: *The Annals of Statistics* 40.3.

Einmahl, John H. J. and Johan Segers (2009). "Maximum Empirical Likelihood Estimation of the Spectral Measure of an Extreme-Value Distribution". In: *The Annals of Statistics* 37 (5B), pp. 2953–2989.

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.

Engelke, Sebastian and Adrien S. Hitz (2019). *Graphical Models for Extremes*. URL: http://arxiv.org/abs/1812.01734 (visited on 11/20/2022). Pre-published.

Engelke, Sebastian and Jevgenijs Ivanovs (2021). "Sparse Structures for Multivariate Extremes". In: *Annual Review of Statistics and Its Application* 8.1, pp. 241–270.

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.

Fix, Miranda J., Daniel S. Cooley, and Emeric Thibaud (2021). "Simultaneous Autoregressive Models for Spatial Extremes". In: *Environmetrics* 32.2.

Fomichov, V and J Ivanovs (2023). "Spherical Clustering in Detection of Groups of Concomitant Extremes". In: *Biometrika* 110.1, pp. 135–153.

Fougères, Anne-Laure, Cécile Mercadier, and John P. Nolan (2013). "Dense Classes of Multivariate Extreme Value Distributions". In: *Journal of Multivariate Analysis* 116, pp. 109–129.

Galambos, Janos (1975). "Order Statistics of Samples from Multivariate Distributions". In: *Journal of the American Statistical Association* 70 (351a), pp. 674–680.

Gissibl, Nadine and Claudia Klüppelberg (2018). "Max-linear models on directed acyclic graphs". In: *Bernoulli* 24 (4A).

Gissibl, Nadine, Claudia Klüppelberg, and Steffen Lauritzen (2019). "Identifiability and Estimation of Recursive Max-Linear Models".

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.

Gong, Yan et al. (2024). "Partial Tail-Correlation Coefficient Applied to Extremal-Network Learning". In: *Technometrics* 66.3, pp. 331–346.

Gudendorf, Gordon and Johan Segers (2010). "Extreme-Value Copulas". In: *Copula Theory and Its Applications*. Ed. by Piotr Jaworski et al. Vol. 198. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 127–145.

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

Huang, Whitney K. et al. (2019). "New Exploratory Tools for Extremal Dependence: $$\chi $$ Networks and Annual Extremal Networks". In: *Journal of Agricultural, Biological and Environmental Statistics* 24.3, pp. 484–501.

Huser, R. and A. C. Davison (2013). "Composite Likelihood Estimation for the Brown-Resnick Process". In: *Biometrika* 100.2, pp. 511–518.

Huser, Raphaël, Anthony C. Davison, and Marc G. Genton (2016). "Likelihood Estimators for Multivariate Extremes". In: *Extremes* 19.1, pp. 79–103.

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.

Janßen, Anja, Sebastian Neblung, and Stilian Stoev (2023). "Tail-Dependence, Exceedance Sets, and Metric Embeddings". In: *Extremes* 26.4, pp. 747–785.

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

Jessen, Anders Hedegaard and Thomas Mikosch (2006). "Regularly Varying Functions". In: *Publications de L'institut Mathematique* 80.94, pp. 171–192.

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.

Joe, Harry (1990). "Families of Min-Stable Multivariate Exponential and Multivariate Extreme Value Distributions". In: *Statistics & Probability Letters* 9.1, pp. 75–81.

Kaufman, Leonard and Peter J. Rousseeuw (1990). *Finding Groups in Data*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Kiriliouk, Anna (2020). "Hypothesis Testing for Tail Dependence Parameters on the Boundary of the Parameter Space". In: *Econometrics and Statistics* 16, pp. 121–135.

Kiriliouk, Anna and Philippe Naveau (2020). "Climate Extreme Event Attribution Using Multivariate Peaks-over-Thresholds Modeling and Counterfactual Theory". In: *The Annals of Applied Statistics* 14.3.

Kiriliouk, Anna and Chen Zhou (2022). *Estimating Probabilities of Multivariate Failure Sets Based on Pairwise Tail Dependence Coefficients*. URL: http://arxiv.org/abs/2210.12618 (visited on 06/13/2023). preprint.

Klüppelberg, Claudia and Mario Krali (2021). "Estimating an Extreme Bayesian Network via Scalings". In: *Journal of Multivariate Analysis* 181, p. 104672.

Krali, Mario (2018). "Causality and Estimation of Multivariate Extremes on Directed Acyclic Graphs". MA thesis. Munich: Technische Universität München.

Larsson, Martin and Sidney Resnick (2012). "Extremal Dependence Measure and Extremogram: The Regularly Varying Case". In: *Extremes* 15.2, pp. 231–256.

Lee, Jeongjin and Daniel Cooley (2023). *Partial Tail Correlation for Extremes*. URL: http://arxiv.org/abs/2210.02048 (visited on 10/19/2023). preprint.

Lehtomaa, Jaakko and Sidney Resnick (2020). "Asymptotic Independence and Support Detection Techniques for Heavy-Tailed Multivariate Data". In: *Insurance: Mathematics and Economics* 93, pp. 262–277.

Liu, Jun and Jieping Ye (2009). "Efficient Euclidean Projections in Linear Time". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09: The 26th Annual International Conference on Machine Learning Held in Conjunction with the 2007 International Conference on Inductive Logic Programming. Montreal Quebec Canada: ACM, pp. 657–664.

Medina, Marco Avella, Richard A. Davis, and Gennady Samorodnitsky (2021). *Spectral Learning of Multivariate Extremes*. URL: http://arxiv.org/abs/2111.07799 (visited on 07/25/2022). Pre-published.

Mestre, Xavier (2008). "Improved Estimation of Eigenvalues and Eigenvectors of Covariance Matrices Using Their Sample Estimates". In: *IEEE Transactions on Information Theory* 54.11, pp. 5113–5129.

Meyer, Nicolas and Olivier Wintenberger (2020). "Detection of Extremal Directions via Euclidean Projections". In: p. 41.

– (2023). "Multivariate Sparse Clustering for Extremes". In: *Journal of the American Statistical Association*, pp. 1–12.

Mhatre, Nehali and Daniel Cooley (2021). "Transformed-Linear Models for Time Series Extremes".

Oesting, Marco, Martin Schlather, and Petra Friederichs (2017). "Statistical Post-Processing of Forecasts for Extremes Using Bivariate Brown-Resnick Processes with an Application to Wind Gusts". In: *Extremes* 20.2, pp. 309–332.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). "Geometric Approach to Statistical Analysis on the Simplex". In: *Stochastic Environmental Research and Risk Assessment* 15.5, pp. 384–398.

Reiss, Rolf-Dieter and Michael Thomas (2007). *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*. Third Edition. SpringerLink Bücher. Basel: Birkhäuser Verlag AG. 511 pp.

Resnick, Sidney (2004). "The Extremal Dependence Measure and Asymptotic Independence". In: *Stochastic Models* 20.2, pp. 205–227.

– (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering. New York, N.Y: Springer. 404 pp.

Richards, Jordan et al. (2024). "Modern Extreme Value Statistics for Utopian Extremes. EVA (2023) Conference Data Challenge: Team Yalla". In: *Extremes*.

Rohrbeck, Christian and Daniel Cooley (2023). "Simulating Flood Event Sets Using Extremal Principal Components". In: *The Annals of Applied Statistics* 17.2.

Russell, Brook T. and Paul Hogan (2018). "Analyzing Dependence Matrices to Investigate Relationships between National Football League Combine Event Performances". In: *Journal of Quantitative Analysis in Sports* 14.4, pp. 201–212.

Schellander, Harald and Tobias Hell (2018). "Modeling Snow Depth Extremes in Austria". In: *Natural Hazards* 94.3, pp. 1367–1389.

Semadeni, Claudio Andri (2020). "Inference on the Angular Distribution of Extremes". PhD thesis. École Polytechnique Fédérale de Lausanne.

Shyamalkumar, Nariankadu D. and Siyang Tao (2020). "On Tail Dependence Matrices: The Realization Problem for Parametric Families". In: *Extremes* 23.2, pp. 245–285.

Simpson, E S, J L Wadsworth, and J A Tawn (2020). "Determining the Dependence Structure of Multivariate Extremes". In: *Biometrika* 107.3, pp. 513–532.

Smith, R L, J A Tawn, and H K Yuen (1990). "Statistics of Multivariate Extremes". In: *International Statistical Review* 58.1, pp. 47–58.

Szemkus, Svenja and Petra Friederichs (2024). "Spatial Patterns and Indices for Heat Waves and Droughts over Europe Using a Decomposition of Extremal Dependency". In: *Advances in Statistical Climatology, Meteorology and Oceanography* 10.1, pp. 29–49.

Tawn, Jonathan A (1990). "Modelling Multivariate Extreme Value Distributions". In: *Biometrika* 77.2, pp. 245–253.

Tran, Ngoc Mai, Johannes Buck, and Claudia Klüppelberg (2021). "Causal Discovery of a River Network from Its Extremes".

Yuen, Robert and Stilian Stoev (2014a). "CRPS M-estimation for Max-Stable Models". In: *Extremes* 17.3, pp. 387–410.

– (2014b). "Upper Bounds on Value-at-Risk for the Maximum Portfolio Loss". In: *Extremes* 17.4, pp. 585–614.

A Properties of the TPDM

A.1 Equivalence of TPDM definitions

We aim to shed light on this matter by showing in the bivariate setting that the TPDM (with respect to some $\alpha \geq 1$) is independent of α. The following lemma helps us achieve this: it gives the formula for transforming between angular densities defined with different α values.

Lemma A.1. *Suppose* $\boldsymbol{X} = (X_i, X_j) \in \mathcal{RV}^2_+(\alpha)$ *for some* $\alpha \geq 1$. *Let* H_α *denote the normalised angular measure with respect to* $\|\cdot\|_\alpha$ *and* $h_\alpha : \mathbb{S}_{+(\alpha)} \to \mathbb{R}_+$ *the corresponding angular density (assuming it exists). Moreover, we define*

$$\tilde{h}_\alpha : [0,1] \to \mathbb{R}_+, \qquad \theta \mapsto h_\alpha\left(\left(\theta, (1-\theta^\alpha)^{1/\alpha}\right)\right).$$

Then

$$\tilde{h}_\alpha(\theta) = \alpha\theta^{\alpha-1}\tilde{h}_1(\theta^\alpha). \tag{A.1}$$

Proof. The proof generalises the procedure described in Section 3.2 of the Supplementary Material of Fix et al. (2021). First, we transform from L_1 polar coordinates $(r, \boldsymbol{\theta})$ to Cartesian coordinates $\boldsymbol{z} = (z_i, z_j) = (r\theta_i, r\theta_j)$. The Jacobian of the transformation is $\|\boldsymbol{z}\|_1^{-1}$ (CITE Prop 1in Cooley et al 2012). Using (2.30) with $\alpha = 1$ and $H_1(\mathrm{d}\boldsymbol{\theta}) = h_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$,

$$\begin{aligned}
\nu(\mathrm{d}r \times \mathrm{d}\boldsymbol{\theta}) &= r^{-2}h_1(\boldsymbol{\theta})\,\mathrm{d}r\,\mathrm{d}\boldsymbol{\theta} \\
&= \|\boldsymbol{z}\|_1^{-2}h_1(\boldsymbol{z}/\|\boldsymbol{z}\|_1)\|\boldsymbol{z}\|_1^{-1}\mathrm{d}\boldsymbol{z} \\
&= \|\boldsymbol{z}\|_1^{-3}h_1(\boldsymbol{z}/\|\boldsymbol{z}\|_1)\mathrm{d}\boldsymbol{z} \\
&= \nu(\mathrm{d}\boldsymbol{z}).
\end{aligned}$$

Next, we transform from tail index $\alpha = 1$ to arbitrary α. Let $\boldsymbol{y} = (y_i, y_j) = (z_i^{1/\alpha}, z_j^{1/\alpha})$. The Jacobian of this transformation is $\alpha^2 y_i^{\alpha-1} y_j^{\alpha-1}$. Note that $\|\boldsymbol{z}\|_1 = y_i^\alpha + y_j^\alpha = \|\boldsymbol{y}\|_\alpha^\alpha$.

$$\nu(\boldsymbol{z}) = [\|\boldsymbol{y}\|_\alpha^\alpha]^{-3} \, h_1\left(\frac{y_i^\alpha}{\|\boldsymbol{y}\|_\alpha^\alpha}, \frac{y_j^\alpha}{\|\boldsymbol{y}\|_\alpha^\alpha}\right) \alpha^2 y_i^{\alpha-1} y_j^{\alpha-1} \mathrm{d}\boldsymbol{y} = \nu(\mathrm{d}\boldsymbol{y}).$$

Finally, we transform to L_α polar coordinates $(s, \boldsymbol{\phi})$ with $s = \|\boldsymbol{y}\|_\alpha$ and $\boldsymbol{\phi} = (\phi_i, \phi_j) = \boldsymbol{y}/s$. By (CITE Lemma 1.1 in Song and Gupta (1997)), the Jacobian is $s(1-\phi_i^\alpha)^{(1-\alpha)/a} = s\phi_j^{1-\alpha}$. We now have

$$\begin{aligned}
\nu(\mathrm{d}\boldsymbol{y}) &= [s^\alpha]^{-3} \, h_1\left(\phi_i^\alpha, \phi_j^\alpha\right) \alpha^2 (s\phi_i)^{\alpha-1} (s\phi_j)^{\alpha-1} s\phi_j^{1-\alpha} \, \mathrm{d}s \, \mathrm{d}\boldsymbol{\phi} \\
&= \alpha s^{-\alpha-1} \alpha \phi_i^{\alpha-1} h_1\left(\phi_i^\alpha, \phi_j^\alpha\right) \, \mathrm{d}s \, \mathrm{d}\boldsymbol{\phi} \\
&= \alpha s^{-\alpha-1} h_\alpha(\boldsymbol{\phi}) \, \mathrm{d}s \, \mathrm{d}\boldsymbol{\phi} \\
&= \nu(\mathrm{d}s \times \mathrm{d}\boldsymbol{\phi}),
\end{aligned}$$

where $h_\alpha(\boldsymbol{\phi}) := \alpha \phi_i^{\alpha-1} h_1\left(\phi_i^\alpha, \phi_j^\alpha\right)$. The final step is to compute $\tilde{h}_\alpha$ by projecting the density h_α, which lives on $\mathbb{S}^1_{+(\alpha)}$, down to $[0, 1]$. Writing $\boldsymbol{\phi}$ as $(\phi, (1-\phi^\alpha)^{1/\alpha})$ gives

$$\tilde{h}_\alpha(\phi) = h_\alpha\left(\left(\phi, (1-\phi^\alpha)^{1/\alpha}\right)\right) = \alpha \phi^{\alpha-1} h_1\left((\phi^\alpha, 1-\phi^\alpha)\right) = \alpha \phi^{\alpha-1} \tilde{h}_1(\phi^\alpha).$$

□

In the trivial case $\alpha = 1$ the formula reduces to $\tilde{h}_1(\theta) = \tilde{h}_1(\theta)$, as one would hope. Setting $\alpha = 2$ yields $\tilde{h}_2(\theta) = 2\theta \tilde{h}_1(\theta^2)$, which matches the formula gives in Fix et al. (2021). Note that $\tilde{h}_\alpha$ is well-defined (i.e. is a normalised density), since

$$\int_0^1 \tilde{h}_\alpha(\theta) \, \mathrm{d}\theta = \int_0^1 \alpha \theta^{\alpha-1} \tilde{h}_1(\theta^\alpha) \, \mathrm{d}\theta = \int_0^1 \tilde{h}_1(\phi) \, \mathrm{d}\phi = 1.$$

We now apply the transformation formula to express the TPDM for any $\alpha \geq 1$ in terms of the angular density $\tilde{h}_1$.

Proposition A.1. *Using the notation of Lemma A.1, the off-diagonal entry in the TPDM of* $\boldsymbol{X}$ *is*

$$\sigma_{ij} = m \int_0^1 \sqrt{u(1-u)} \, \tilde{h}_1(u) \, \mathrm{d}\phi. \tag{A.2}$$

Proof. The relation between the normalised measure H_α and the measure H in Definition 2.13 is $H_\alpha = m^{-1}H$, where m is the mass of H. Therefore, (2.52) can be equivalently restated as

$$\sigma_{ij} = m \int_{\mathbb{S}_{+(\alpha)}} \theta_i^{\alpha/2}\theta_j^{\alpha/2} \,\mathrm{d}H_\alpha(\boldsymbol{\theta})$$

Rewriting this in terms of the angular density and re-parametrising yields

$$\begin{aligned}\sigma_{ij} &= m \int_{\mathbb{S}_{+(\alpha)}} \theta_i^{\alpha/2}\theta_j^{\alpha/2} h_\alpha(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} \\ &= m \int_{\mathbb{S}_{+(\alpha)}} \theta_i^{\alpha/2}[(1-\theta_i^\alpha)^{1/\alpha}]^{\alpha/2} h_\alpha(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} \\ &= m \int_0^1 \theta^{\alpha/2}(1-\theta^\alpha)^{1/2} \tilde{h}_\alpha(\theta) \,\mathrm{d}\theta.\end{aligned}$$

Finally, we apply Lemma A.1 and substitute $u = \theta^\alpha$ to obtain the final result

$$\sigma_{ij} = m \int_0^1 \theta^{\alpha/2}(1-\theta^\alpha)^{1/2}\alpha\theta^{\alpha-1}\tilde{h}_1(\theta^\alpha) \,\mathrm{d}\theta = m \int_0^1 \sqrt{u(1-u)}\, \tilde{h}_1(u) \,\mathrm{d}\phi.$$

□

Extra things to find a place for:

Symmetric logistic angular density:

$$\tilde{h}_1(\theta;\gamma) = \frac{1-\gamma}{2\gamma}[\theta(1-\theta)]^{\frac{1}{\gamma}-2}[\theta^{1/\gamma} + (1-\theta)^{1/\gamma}]^{\gamma-2}$$

Hüsler-Reiss angular density:

$$\tilde{h}_1(\theta;\lambda) = \frac{\exp(-\lambda/4)}{4\lambda[\theta(1-\theta)]^{3/2}} \phi\left(\frac{1}{2\lambda}\log\left(\frac{\theta}{1-\theta}\right)\right)$$

A.2 Formula for the asymptotic variance ν_{ij}^2

Adopting the notation of Proposition A.1, the asymptotic variance can be expressed in terms of the angular density $\tilde{h}_1$ of (X,X_j). Using $\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$, we have

$$\nu_{ij}^2 = m^2 \int_{\mathbb{S}_{+(\alpha)}^{d-1}} (\theta_i\theta_j)^\alpha \,\mathrm{d}H_\alpha(\boldsymbol{\theta}) - \sigma_{ij}^2 = m^2 \int_0^1 \theta^\alpha(1-\theta^\alpha)\tilde{h}_\alpha(\theta) \,\mathrm{d}\theta - \sigma_{ij}^2.$$

Substituting $u = \theta^\alpha$ and using Proposition A.1 gives the final expression

$$\nu_{ij}^2 = m^2 \int_0^1 u(1-u)\,\tilde{h}_1(u)\,\mathrm{d}u - \left[m \int_0^1 \sqrt{u(1-u)}\,\tilde{h}_1(u)\,\mathrm{d}u\right]^2. \tag{A.3}$$

The asymptotic distribution of $\hat{\sigma}_{ij}$ does not depend on α.

A.3 Proof of Proposition 2.6

Proof. We follow the proof of Theorem 5.23 in CITE Krali Thesis but adapt it to the general α case. By the Cramér-Wold device (CITE), it is sufficient to show asymptotic normality of $\sqrt{k}\boldsymbol{\beta}^T(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})$ for all $\boldsymbol{\beta} \in \mathbb{R}^{\binom{d}{2}}$. For convenience, the components of $\boldsymbol{\beta}$ are indexed to match the sub-indices of $\boldsymbol{\sigma}$. Then

$$\boldsymbol{\beta}^T\boldsymbol{\sigma} = \sum_{i=1}^d \sum_{j=i}^d \beta_{ij}\sigma_{ij} = \mathbb{E}_{\boldsymbol{\Theta}\sim H}\left[\sum_{i=1}^d \sum_{j=i}^d \beta_{ij}\Theta_i^{\alpha/2}\Theta_j^{\alpha/2}\right] =: \mathbb{E}_{\boldsymbol{\Theta}\sim H}[g(\boldsymbol{\Theta};\boldsymbol{\beta})],$$

where

$$g(\boldsymbol{\theta};\boldsymbol{\beta}) := \sum_{i=1}^d \sum_{j=i}^d \beta_{ij}\theta_i^{\alpha/2}\theta_j^{\alpha/2}$$

The corresponding empirical estimator is

$$\hat{\mathbb{E}}_{\boldsymbol{\Theta}\sim H}[g(\boldsymbol{\Theta};\boldsymbol{\beta})] = \frac{m}{k}\sum_{l=1}^k \sum_{i=1}^d \sum_{j=i}^d \beta_{ij}\Theta_{(l),i}^{\alpha/2}\Theta_{(l),j}^{\alpha/2} = \sum_{i=1}^d \sum_{j=i}^d \beta_{ij}\left(\frac{m}{k}\sum_{l=1}^k \Theta_{(l),i}^{\alpha/2}\Theta_{(l),j}^{\alpha/2}\right) = \boldsymbol{\beta}^T\hat{\boldsymbol{\sigma}}.$$

Noting that $g(\cdot\,;\boldsymbol{\beta})$ is continuous and applying **??**, we have

$$\sqrt{k}\boldsymbol{\beta}^T(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) = \sqrt{k}\left(\hat{\mathbb{E}}_{\boldsymbol{\Theta}\sim H}[g(\boldsymbol{\Theta};\boldsymbol{\beta})] - \mathbb{E}_{\boldsymbol{\Theta}\sim H}[g(\boldsymbol{\Theta};\boldsymbol{\beta})]\right) \to N(0, v(\boldsymbol{\beta})).$$

where $v(\boldsymbol{\beta}) := \mathrm{Var}_{\boldsymbol{\Theta}\sim H}(g(\boldsymbol{\Theta};\boldsymbol{\beta}))$. The asymptotic normality of $\hat{\boldsymbol{\sigma}}$ follows by the Cramér-Wold device. The diagonal elements of the covariance matrix V are as in Proposition 2.5. The off-diagonal entries are given by

$$\begin{aligned}
2\mathrm{Cov}\left(\sqrt{k}(\hat{\sigma}_{ij} - \sigma_{ij}), \sqrt{k}(\hat{\sigma}_{lm} - \sigma_{lm})\right) &= 2k\,\mathrm{Cov}(\hat{\sigma}_{ij}, \hat{\sigma}_{lm}) \\
&= k\left[\mathrm{Var}(\hat{\sigma}_{ij} + \hat{\sigma}_{lm}) - \mathrm{Var}(\hat{\sigma}_{ij}) - \mathrm{Var}(\hat{\sigma}_{lm})\right] \\
&\to \mathrm{Var}_{\boldsymbol{\Theta}\sim H}(\Theta_i^{\alpha/2}\Theta_j^{\alpha/2} + \Theta_l^{\alpha/2}\Theta_m^{\alpha/2}) - \nu_{ij}^2 - \nu_{lm}^2.
\end{aligned}$$

□

A.4 Derivation of the asymptotic covariance matrix V under the max-linear model

Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ is max-linear with q factors and parameter matrix A. Then, for any $i, j = 1, \ldots, d$, we have $\sigma_{ij} = \sum_{l=1}^q a_{il}^{\alpha/2} a_{jl}^{\alpha/2}$ and

$$\nu_{ij}^2 = d \int_{\mathbb{S}_{+(\alpha)}^{d-1}} (\theta_i \theta_j)^\alpha \, \mathrm{d}H(\boldsymbol{\theta}) - \sigma_{ij}^2 = d \sum_{s=1}^q \|\boldsymbol{a}_s\|_\alpha^\alpha \left(\frac{a_{is} a_{js}}{\|\boldsymbol{a}_s\|_\alpha^2} \right)^\alpha - \sigma_{ij}^2 = d \sum_{s=1}^q \frac{(a_{is} a_{js})^\alpha}{\|\boldsymbol{a}_s\|_\alpha^\alpha} - \sigma_{ij}^2.$$

For any pair of upper-triangular index pairs (i, j) and (l, m), we have

$$\begin{aligned}
&\mathrm{Var}_{\boldsymbol{\Theta} \sim H}(\Theta_i^{\alpha/2} \Theta_j^{\alpha/2} + \Theta_l^{\alpha/2} \Theta_m^{\alpha/2}) \\
&\quad = d \int_{\mathbb{S}_{+(\alpha)}^{d-1}} \left[(\theta_i \theta_j)^\alpha + 2(\theta_i \theta_j \theta_l \theta_m)^{\alpha/2} + (\theta_l \theta_m)^\alpha \right] \mathrm{d}H(\boldsymbol{\theta}) - \left[\sigma_{ij} + \sigma_{lm} \right]^2 \\
&\quad = d \sum_{s=1}^q \frac{(a_{is} a_{js})^\alpha + 2(a_{is} a_{js} a_{ls} a_{ms})^{\alpha/2} + (a_{ls} a_{ms})^\alpha}{\|\boldsymbol{a}_s\|_\alpha^\alpha} - \left[\sigma_{ij} + \sigma_{lm} \right]^2 \\
&\quad = \nu_{ij}^2 + \nu_{lm}^2 + d \sum_{s=1}^q \frac{2(a_{is} a_{js} a_{ls} a_{ms})^{\alpha/2}}{\|\boldsymbol{a}_s\|_\alpha^\alpha} - 2\sigma_{ij}\sigma_{lm}
\end{aligned}$$

and therefore

$$2\rho_{ij,lm} = d \sum_{s=1}^q \frac{2(a_{is} a_{js} a_{ls} a_{ms})^{\alpha/2}}{\|\boldsymbol{a}_s\|_\alpha^\alpha} - 2\sigma_{ij}\sigma_{lm}.$$

The expressions for ν_{ij}^2 and $\rho_{ij,lm}$ can be summarised as

$$v_{ij,lm} = d \sum_{s=1}^q \frac{(a_{is} a_{js} a_{ls} a_{ms})^{\alpha/2}}{\|\boldsymbol{a}_s\|_\alpha^\alpha} - \sigma_{ij}\sigma_{lm}. \tag{A.4}$$