Extensions and Applications of the Tail Pairwise Dependence Matrix

Matthew Pawley

October 14, 2024

Table of contents

List of Figures

List of Tables

Preface

Draft thesis of Matthew Pawley, created on October 14, 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 and identically distributed, continuous random variables with distribution function F. For $n \geq 1$, define

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

The distribution of M_n is given by

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

In practice, this result is not particularly useful, since F is usually unknown. Instead, we leverage asymptotic theory to study the limiting behaviour of F^n as $n \to \infty$. The asymptotic distribution of M_n is degenerate, since $M_n \xrightarrow{p} x_F$, the (possibly infinite) upper end-point of F. The Extremal Types Theorem states that, after suitable rescaling, there are three classes of non-degenerate asymptotic distribution.

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

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

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: $\xi > 0$ corresponds to the heavy-tailed Fréchet class, $\xi = 0$ (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. Let $x_1, \ldots, x_n$ denote independent observations of $X_1, \ldots, X_n$. The data are then partitioned into finite blocks of size m. Provided m is sufficiently large, Theorem 2.1 implies that the maximum observation in each block is approximately GEV distributed. Estimates for the GEV parameters are obtained from the set of block-wise maxima, e.g. by maximum likelihood inference. The choice of block size is critical and involves managing a bias-variance trade-off. If m is too small, then the asymptotic approximation may not be valid. If m is too large, then the number of blocks (and therefore block-wise maxima) will be small, resulting in high variances in the estimates.

2.1.2 Threshold exceedances and the generalised Pareto distribution (GPD)

The block maxima procedure is considered wasteful, because it fails to exploit all the available information. Specifically, observations that are ‘extreme’ but not block maxima are discarded, even though they can be informative for the tail behaviour. This motivates the alternative but intimately related peaks-over-threshold method. 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 $\tilde{\sigma} = \sigma + \xi(u - \mu)$. The limiting conditional distribution is called the generalised Pareto distribution (GPD). Given observations $x_1, \ldots, x_n$, it suggests itself to choose a high threshold u, and assume that exceedances of the threshold are approximately GPD distributed. The GPD parameters can then be estimated by likelihood or Bayesian inference procedures. Threshold selection is subject to similar considerations as for the block size. If the threshold is too low then the GPD model is not valid, leading to a bias in the fitted model. If the threshold is too high, then the uncertainty in the estimated model parameters will be unnecessarily high. Many diagnostics and methodologies are proposed in the literature to aid with this choice.

2.1.3 Regular variation

Definition 2.1. A function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying with index $\alpha \in \mathbb{R}$ if, for all $x > 0$,

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\alpha. \tag{2.5}$$

In the case $\alpha = 0$, f is called slowly-varying. This notion is extended to random variables by treating the distributional tail as the function of interest.

Definition 2.2. 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

$$\mathbb{P}(X > x) = x^{-\alpha} L(x) \tag{2.6}$$

for some slowly-varying function L (Jessen and Mikosch 2006). This says that 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. Moreover, (2.6) means that regularly varying distributions are asymptotically scale invariant, in the sense that for all $\lambda > 0$,

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

This asymptotic homogeneity explains why regular variation is ubiquitous in extreme value theory.

2.1.4 Non-stationary extremes

To do.

2.2 Multivariate extreme value theory

2.2.1 Componentwise maxima

Let $\boldsymbol{X} = (X_1, \dots, X_d)$ be a d-dimensional random vector with unknown joint distribution function F. That is, for any $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$,

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

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ be a sequence of independent copies of $\boldsymbol{X}$. In the multivariate setting, the notion of a 'maximum' becomes subjective, since $\mathbb{R}^d$ is not an ordered set. One possibility is to define the maximum component-wise as

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

Analgously to the univariate case, 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

$$\mathbb{P}\left(\frac{\boldsymbol{M}_n - \boldsymbol{b}_n}{\boldsymbol{a}_n} \leq \boldsymbol{x} \right) \xrightarrow{d} G(\boldsymbol{x}), \qquad (n \to \infty). \tag{2.7}$$

By application of Theorem 2.1 to the marginal components, one can show that the margins of G follow a univariate GEV distribution. Unlike the univariate case, the full (joint) limit distribution G does not admit a parametric representation. To study the properties of G, it is usual to standardise to common margins.

2.2.2 Copulae and marginal standardisation

In multivariate statistics, copula theory provides a way to divide the modelling process into two steps: modelling the margins and modelling the dependence between the variables.

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*

$$F(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_d(x_d)\right). \tag{2.8}$$

The copula C characterises the dependence structure of the variables. It represents the distribution function of $\boldsymbol{X}$ after transforming to standard uniform margins. Uniform margins are a standard choice in multivariate statistics, but one could easily conceive of copulae defined with alternative marginal distributions. In extreme value theory, it is common to work with Fréchet, exponential, or Gumbel margins. The different choices accentuate particular features of the extreme values. For example, heavy-tailed Fréchet margins will highlight the most extreme values. Light-tailed Gumbel or exponential margins are often preferred for conditional extremes modelling (CITE Heffernan and Tawn).

There are broadly two ways of performing the preliminary 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 distribution $F_\star$ via by the probability integral transform:

$$X_i \mapsto F_\star^{-1}(F_i(X_i)) \sim F_\star, \qquad (i = 1, \ldots, d). \tag{2.9}$$

In applications, the marginal distributions are usually unknown. Then an estimate $\hat{F}_i$ replaces F_i in (2.9). This is typically the empirical CDF (non-parametric) possibly with GPD tails above a high threshold (semi-parametric). Uncertainty arising from the empirical marginal standardisation step will be neglected in this thesis.

2.2.3 The exponent measure and angular measure

Suppose $\boldsymbol{X}$ is on unit Fréchet margins, so that for $i = 1, \ldots, d$,

$$\mathbb{P}(X_i < x) = \exp(-1/x), \qquad (x > 0). \tag{2.10}$$

This corresponds to a GEV distribution (2.3) with $\mu = \sigma = \xi = 1$. Then the distribution G in (2.7) may be expressed as

$$G(\boldsymbol{x}) = \exp(-V(\boldsymbol{x})), \tag{2.11}$$

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

$$V(\boldsymbol{x}) = d \int_{\mathbb{S}^{d-1}_{+(1)}} \bigvee_{i=1}^{d} \left(\frac{\theta_i}{x_i} \right) \mathrm{d}H(\boldsymbol{\theta}), \tag{2.12}$$

where

$$\mathbb{S}^{d-1}_{+(1)} := \{\boldsymbol{x} \in \mathbb{R}^d_+ : \|\boldsymbol{x}\|_1 = 1\} \tag{2.13}$$

is the ℓ_1-simplex in $\mathbb{R}^d$ and the angular measure H is a probability measure that satisfies the moment constraints

$$\int_{\mathbb{S}^{d-1}_{+(1)}} \theta_i \, \mathrm{d}H(\boldsymbol{\theta}) = 1/d, \qquad (i = 1, \dots, d). \tag{2.14}$$

These constraints stem from the fixing of the margins. Functions G satisfying (2.11) 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

$$h\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1} \right) = -\frac{\|\boldsymbol{x}\|_1^{d+1}}{d} \frac{\partial^d}{\partial x_1 \cdots \partial x_d} V(\boldsymbol{x}). \tag{2.15}$$

Models for G are typically defined by specifying a parametric form for V or H.

2.2.4 Parametric multivariate extreme value models

The class of valid dependence structures is in direct correspondence to the class of valid measures H, which is infinite-dimensional. This is a significant impediment to performing statistics: efficient estimation via likelihood inference, hypothesis testing, and inclusion of covariates become unavailable. To circumvent this, one may postulate a parametric sub-family that generates a wide class of valid dependence structures. Many models of this kind have been proposed in the literature; a detailed review can be found in Gudendorf and Segers (2010). The price paid is that working within a sub-family of the general class runs the risk of model misspecification. Generating valid models is challenging, owing to the awkward moment constraints. This results in distribution functions and parameter constraints that are often cumbersome and unwieldy. Moreover, striking the balance between flexibility and parsimony becomes difficult 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 that will feature regularly as data-generating processes for our numerical experiments.

2.2.4.1 Logistic-type models

The simplest model is the symmetric logistic distribution (CITE Gumbel 1960).

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

$$V(\boldsymbol{x}) = \left(\sum_{i=1}^{d} x_i^{-1/\gamma} \right)^{\gamma}, \qquad \gamma \in (0,1]. \tag{2.16}$$

The single dependence parameter $\gamma \in (0,1]$ characterises the strength of (tail) association between all variables. The variables are independent when $\gamma = 1$. As $\gamma \to 0$ they approach complete dependence. The distribution function is invariant under coordinate permutation, meaning the variables are exchangeable. A flexible extension is the asymmetric logistic model of Tawn (1990). It permits greater control over the dependence structure at the expense of a greater number of parameters.

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

$$V(\boldsymbol{x}) = \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset} \left[\sum_{i \in \beta} \left(\frac{\theta_{i,\beta}}{x_i} \right)^{1/\gamma_\beta} \right]^{\gamma_\beta}, \qquad \begin{cases} \gamma_\beta \in (0,1], & \\ \theta_{i,\beta} \in [0,1], & \text{if } i \in \beta, \\ \theta_{i,\beta} = 0, & \text{if } i \notin \beta, \\ \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset} \theta_{i,\beta} = 1, & \end{cases} \tag{2.17}$$

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

The parameters γ_β control the dependence between among $\{X_i : i \in \beta\}$ in a similar way to the symmetric logistic model. The asymmetry parameters $\boldsymbol{\theta}_\beta = (\theta_{i,\beta} : i \in \beta)$ *explain their interpretation...* . Further models can be generated by 'inverting' the logistic and asymmetric models. This yields the negative logistic model (CITE Galambos 1975) and the negative asymmetric logistic model (Joe 1990), respectively.

Definition 2.5. The exponent measure of a random vector $\boldsymbol{X} = (X_1,\dots,X_d)$ following the negative symmetric logistic distribution is

$$V(\boldsymbol{x}) = \sum_{\beta \in \mathcal{P}(\{1,\dots,d\})\setminus\emptyset} (-1)^{|\beta|+1} \left(\sum_{i\in\beta} x_i^\gamma \right)^{-1/\gamma}, \qquad \gamma > 0. \tag{2.18}$$

Definition 2.6. The exponent measure of a random vector $\boldsymbol{X} = (X_1,\dots,X_d)$ following the negative asymmetric logistic distribution is

$$V(\boldsymbol{x}) = \sum_{\beta \in \mathcal{P}(\{1,\dots,d\})\setminus\emptyset} (-1)^{|\beta|+1} \left(\sum_{i\in\beta} x_i^\gamma \right)^{-1/\gamma}, \qquad \gamma > 0. \tag{2.19}$$

Definition 2.7. Smith et al. (1990)

2.2.4.2 Brown-Resnick processes and the Hüsler-Reiss distribution

Consider a Brown-Resnick process $\{X(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^2\}$ with semi-variogram

$$\gamma(\boldsymbol{s}, \boldsymbol{s}') = (\|\boldsymbol{s} - \boldsymbol{s}'\|_2/\rho)^\kappa, \qquad \rho > 0, \kappa \in (0, 2]. \tag{2.20}$$

Semi-variograms of the form (2.20) are called fractal semi-variograms. The associated spatial process $X(\boldsymbol{s})$ is stationary and isotropic (**engelke_estimation_2015**). The parameters ρ and κ control the range and smoothness, respectively. *Davison et al. (2012) apply to rainfall data, finding* $1/2 < \kappa < 1$.

Definition 2.8. The bivariate exponent measure of a Brown-Resnick process $X(\boldsymbol{s})$ at sites $\{\boldsymbol{s}_i, \boldsymbol{s}_j\}$ is (R. Huser and Davison 2013)

$$V(x_i, x_j) = \frac{1}{x_i}\Phi\left(\frac{a_{ij}}{2} + \frac{1}{a_{ij}}\log\frac{x_j}{x_i}\right) + \frac{1}{x_j}\Phi\left(\frac{a_{ij}}{2} + \frac{1}{a_{ij}}\log\frac{x_i}{x_j}\right), \tag{2.21}$$

where $x_i = x(\boldsymbol{s}_i)$, $x_j = x(\boldsymbol{s}_j)$, and $a_{ij} = \sqrt{\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)}$.

From (2.21) it is evident that the association between two sites is determined by their spatial proximity, since a_{ij} depends on the spatial locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ only through $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2$. This reflects the stationarity of the underlying spatial process.

The Brown-Resnick process is intimately related to the Hüsler-Reiss distribution of Hüsler and Reiss (1989), which arises as the limit of suitably normalised Gaussian random vectors. The Hüsler-Reiss distribution is of fundamental importance in multivariate extremes; it has been labelled the Gaussian distribution for extremes (**engelke_graphical_2019**). In $d \geq 2$ dimensions the distribution is parametrised by a matrix $\Lambda = (\lambda_{ij}^2)_{1\leq i,j\leq d} \in \mathcal{D}$, where $\mathcal{D} \subset \mathbb{R}_+^{d\times d}$ denotes the space of symmetric, strictly conditionally negative definite matrices

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

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 Λ, then (X_i, X_j), $i \neq j$, is also Hüsler-Reiss distributed with parameter λ_{ij}^2. The dependence between any pairs of components can be controlled by modifying the corresponding $\lambda_{ij} > 0$, subject to the constraint $\Lambda \in \mathcal{D}$. Its relation to the Brown-Resnick process is that the finite-dimensional distribution at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d$ of a Brown-Resnick process is the Hüsler-Reiss distribution with $\Lambda = (\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)/4)_{1\leq i,j\leq d}$ (**engelke_estimation_2015**). Due to this connection, the Hüsler-Reiss distribution is often parametrised by the variogram matrix $\Gamma = 4\Lambda \in \mathcal{D}$ (**engelke_sparse_2021**; **fomichov_spherical_2023**). The bivariate exponent measure of (X_i, X_j) is given by (2.21) with a_{ij} replaced with $2\lambda_{ij}$.

2.2.4.3 The max-linear model

Preamble.

Definition 2.9. Let $A = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_q) \in \mathbb{R}_+^{d\times q}$ for some $q \geq 1$. Assume that $\boldsymbol{a}_j \neq \boldsymbol{0}$ for all $j = 1, \dots, q$ and each row has unit sum, i.e $\sum_{j=1}^q a_{ij} = 1$ for $i = 1, \dots, d$. A random vector $\boldsymbol{X} = (X_1, \dots, X_d)$ with discrete angular measure

$$H(\cdot) = \frac{1}{\sum_{j=1}^q \|\boldsymbol{a}_j\|_1} \sum_{j=1}^q \|\boldsymbol{a}_j\|_1 \delta_{\boldsymbol{a}_j/\|\boldsymbol{a}_j\|_1}(\cdot) \tag{2.22}$$

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

The unit-sum constraint on the rows of A ensures that (2.22) is a valid angular measure since, for any $i = 1, \dots, d$,

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

Due to the row constraints the max-linear model has $d\times(q-1)$ free parameters. Reordering the columns does not alter the angular measure. The factors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_q$ correspond to the possible directions that extremal observations may take, while $\|\boldsymbol{a}_1\|_1, \dots, \|\boldsymbol{a}_q\|_1$ determine their respective weights.

To construct a random vector $\boldsymbol{X} = (X_1, \dots, X_d)$ with angular measure (2.22), we let $Z_1, \dots, Z_q$ be independent unit Fréchet random variables and $\boldsymbol{Z} = (Z_1, \dots, Z_q)$ and set

$$\boldsymbol{X} = A \times_{\max} \boldsymbol{Z} := \left(\bigvee_{j=1}^q a_{1j} Z_j, \dots, \bigvee_{j=1}^q a_{dj} Z_j \right) \tag{2.23}$$

or

$$\boldsymbol{X} = A \otimes \boldsymbol{Z} := \bigoplus_{j=1}^q (\boldsymbol{a}_j \odot Z_j). \tag{2.24}$$

The operations $\oplus$ and $\odot$ relate to the vector space in Cooley and Thibaud (2019); they will be defined explicitly in Section XX. With this construction, $\boldsymbol{X}$ has angular measure (2.22) and unit Fréchet margins (Kiriliouk and Zhou 2022). 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 (2.23) with q factors. The class of angular measures (2.22) 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 an angular measure generated by a max-linear model with sufficiently many factors. This makes the max-linear model a versatile and powerful modelling framework, despite its simplicity.

Formulae for tail events under max-linear model.

2.2.4.4 Sampling from parametric models

The logistic-type, Hüsler-Reiss and max-linear models will be used to generate synthetic data throughout this thesis. The R package `mev` provides functionalities for this purpose. The underlying sampling algorithms are formulated in Dombry et al. (2016). The `rmev` function generates independent realisations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ on unit Fréchet margins from the specified parametric multivariate extreme value model. The function `rmevspec` produces independent observations $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_n$ directly from the angular measure H. Generally, we will be interested in using the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ to learn a model for H. The samples $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_n$ can be used for model validation.

Make figure depicting the dependence structure of SL, HR and max-linear models.

2.2.5 Multivariate regular variation

Multivariate regular variation (MRV) provides an alternative characterisation of the probabilistic structure of multivariate extreme events. Under this framework, the asymptotic joint tail distribution is represented by a homogeneous limit measure. Although MRV can be formulated more generally, we focus on the case where $\boldsymbol{X}$ takes 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 focus on the lows or highs of precipitation records, depending on whether he is assessing drought risk or flood risk. Without loss of generality, and by means of a transformation if necessary, we can define this direction of interest to be 'positive'.

Definition 2.10. We say that $\boldsymbol{X}$ 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

$$n\mathbb{P}(b_n^{-1}\boldsymbol{X} \in \cdot) \xrightarrow{\text{v}} \nu(\cdot), \qquad (n \to \infty), \tag{2.25}$$

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 $s > 0$,

$$\nu(s\,\cdot) = s^{-\alpha}\nu(\cdot). \tag{2.26}$$

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

$$\mathbb{S}_+^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\| = 1\} \tag{2.27}$$

such that

$$n\mathbb{P}((b_n^{-1}R, \boldsymbol{\Theta}) \in \cdot) \xrightarrow{\text{v}} \nu_\alpha \times H(\cdot), \qquad (n \to \infty), \tag{2.28}$$

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 (2.25) and (2.28) are related via

$$\nu(\{\boldsymbol{x} \in \mathbb{E}_0 : \|\boldsymbol{x}\| > s, \boldsymbol{x}/\|\boldsymbol{x}\| \in \cdot\}) = s^{-\alpha}H(\cdot), \qquad \nu(\mathrm{d}r \times \mathrm{d}\boldsymbol{\theta}) = \alpha r^{-\alpha-1}\mathrm{d}r\,\mathrm{d}H(\boldsymbol{\theta}). \tag{2.29}$$

The pseudo-polar formulation (2.28) reveals the attractive feature of MRV. It states that the extremal behaviour of $\boldsymbol{X}$ is fully characterised by two objects. The tail index α represents the index of regular variation of the radial component, thereby governing the heavy-tailedness of $\|\boldsymbol{X}\|$. The angular measure H fully characterises the dependence structure. Crucially, the right-hand side of (2.28) is a product measure. This is often called the radial-angular decomposition This signifies 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. Recalling from Section XX that standard practice is to standardise the margins prior to modelling the dependence structure, this is not constraining. Fixing the marginal distributions determines the index α. In this thesis, we will typically choose Fréchet margins with unit scale and shape parameter $\alpha > 0$, that is

$$\mathbb{P}(X_i < x) = \exp(-x^{-\alpha}), \qquad (x > 0). \tag{2.30}$$

A random vector with α-Fréchet margins (2.30) has tail index α.

The angular measure is unique with respect to a fixed norm $\|\cdot\|$ and lies on the corresponding unit simplex defined in (2.27). The exponent $d-1$ in $\mathbb{S}_+^{d-1}$ references that fact that the simplex is a $(d-1)$-dimensional set embedded in the d-dimensional Euclidean space $\mathbb{R}^d$. In this thesis, we will exclusively use the L_p-norm

$$\|\cdot\|_p : \mathbb{R}^d \to \mathbb{R}, \qquad \|\boldsymbol{x}\|_p = \left(\sum_{i=1}^d x_i^p\right)^{1/p} \tag{2.31}$$

The corresponding simplex will be denoted by

$$\mathbb{S}_{+(p)}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\|_p = 1\}. \tag{2.32}$$

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 (2.28). To see this, note that replacing $\{b_n\}$ by $\{s\cdot b_n\}$ for some $s > 0$ yields a new angular measure $H' = s^{-\alpha}H$ with total mass $m' = s^{-\alpha}m$. We are free to choose whether the scaling information is contained in $\{b_n\}$ or H. Fougères et al. (2013) explain possible reasons for preferring one over the other, but ultimately it is an arbitrary modelling choice. Conventionally H is normalised to be a probability measure, that is $m = 1$. At certain points during this thesis, we might instead specify $\{b_n\}$ and push the scaling information on to H. Irrespective of whether H is normalised or not, we write $\boldsymbol{W} \sim H$ to denote a random vector $\boldsymbol{W}$ whose distribution is the probability measure $m^{-1}H$.

With $\boldsymbol{X}$ standardised to common margins, the centre of mass of H must lie in the simplex interior:

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

The value of μ depends on the choice of norm. If $\|\cdot\| = \|\cdot\|_1$, then $\mu = m/d$, in accordance with (2.14). If $\|\cdot\| = \|\cdot\|_2$, then $m/d \leq \mu \leq m/\sqrt{d}$ (Fomichov and Ivanovs 2023, Lemma 2.1). ### Extremal dependence

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 can be very complex, being subject only to the mean constraints (2.14). For example, the extremal dependence between a meteorological variable measured at two locations may depend on the topography of the spatial domain, the physics of the underlying climatological processes, and the locations' spatial proximity.

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

2.2.6 The tail dependence coefficient

Definition 2.11. 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 $\chi_\beta =: \chi_{ij}$.

If $\chi_{ij} = 0$, then we say that X_i and X_j are asymptotically independent. This means that X_i and X_j cannot take their largest values simultaneously. If $\chi_{ij} \in (0, 1]$, then the variables exhibit asymptotic dependence 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 relation between the tail dependence coefficient and the angular measure is as follows: $\chi_\beta > 0$ if and only if there exists $\beta' \supset \beta$ 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 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 $\boldsymbol{e}_1, \dots, \boldsymbol{e}_d$ denote the canonical basis vectors of $\mathbb{R}^d$. The measure H_1 places all its mass on the coordinate axes. This corresponds to the case of full asymptotic independence, i.e. $\chi_\beta = 0$ for all (non-empty, non-singleton) $\beta \subseteq \{1, \dots, d\}$. On the other hand, a random vector with angular measure H_2 possesses perfect/complete asymptotic dependence and $\chi_\beta > 0$ for all β.

If the bivariate exponent measure $V(x_i, x_j)$ of (X_i, X_j) is known, then the tail dependence coefficient can be computed using $\chi_{ij} = 2 - V(1,1)$.

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

$$\chi_{ij} = 2 - V_{ij}(1,1) = 2 - \left[\left(x_i^{-1/\alpha} + x_j^{-1/\alpha}\right)^\alpha\right] = 2 - 2^\alpha. \tag{2.37}$$

Therefore X_i and X_j approach asymptotic independence when $\alpha = 1$ and exhibit asymptotic dependence when $\alpha \in (0,1)$.

Example 2.2. See Simpson thesis page 18-19.

Example 2.3. Let $\boldsymbol{X} = (X_1, \dots, X_d)$ 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

$$\chi_{ij} = 2 - V_{ij}(1,1) = 2 - 2\Phi\left(\lambda_{ij} + \frac{1}{2\lambda_{ij}} \log 1\right) = 2 - 2\Phi(\lambda_{ij}). \tag{2.38}$$

This concurs with, e.g. Remark 25 in CITE Kabluchko et al. (2009). We note that X_i and X_j are asymptotically dependent for all $\lambda > 0$, with asymptotic independence in the limit as $\lambda \to \infty$.

Example 2.4. Suppose $\boldsymbol{X} = (X_1, X_2)$ is max-linear with parameter matrix $A \in \mathbb{R}_+^{2\times q}$. Then using the angular measure (**??**) and its relation to the exponent measure (2.12), we have

$$\chi_{12} = 2 - V_{12}(1,1) = 2 - 2\int_{\mathbb{S}^1_{+(1)}} (\theta_1 \vee \theta_2)\,\mathrm{d}H(\boldsymbol{\theta}) = 2 - 2\sum_{j=1}^{q}(a_{1j} \vee a_{2j}). \tag{2.39}$$

The bivariate dependence measure χ_{ij} is usually estimated by computing the empirical probabilities $\hat{\chi}_{ij}(u)$ at a sequence of high quantiles u approaching one. An example of the resulting diagnostic plot is provided in Figure 2.1. The underlying data are generated from a symmetric logistic model with $\alpha = 0.5$. The black points represent the empirical estimates $\hat{\chi}_{ij}(u)$ over the range $0.8 \leq u \leq 0.995$, with a 95% confidence interval depicted by the grey region. The true value $\chi_{ij} = 2 - \sqrt{2} \approx 0.59$ (see **??**) is indicated by the red horizontal line. The plot illustrates a clear example of the bias-variance trade-off in relation to the choice of quantile/threshold. This phenomenon is ubiquitous in threshold-based extreme value statistics and will be discussed in more detail in Section ??.

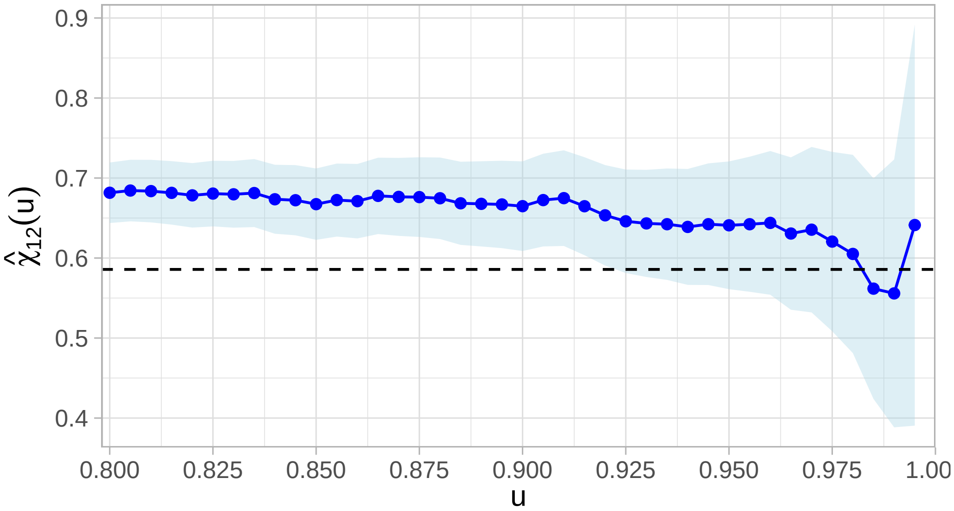

Figure 2.1: Empirical estimates $\hat{\chi}_{12}(u)$ of the tail dependence coefficient for bivariate symmetric logistic data with $\gamma = 0.5$ and $n = 5,000$ observations. The true coefficient $\chi_{12} = 2 - 2^{\gamma} \approx 0.59$ is marked by the dashed line. The shaded region represents the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

Estimation of χ_β for $|\beta| > 2$ is more complicated. Determining the collection of β for which $\chi_\beta > 0$ is equivalent to identifying the support of the angular measure, i.e. which faces of the simplex possess H-mass. *This will be explained later. Sparsity assumption, empirical angular measure only places mass on interior, etc.*

2.2.6.1 Extremal dependence measure

An alternative bivariate summary measure is the extremal dependence measure (EDM). It was originally proposed by Resnick (2004) and later refined by Larsson and Resnick (2012).

Definition 2.12. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ with angular measure H. The EDM between X_i and X_j is

$$\mathrm{EDM}_{ij} := \int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta}) \, \mathrm{d}H(\boldsymbol{\theta}), \qquad f(\boldsymbol{\theta}) = \theta_i \theta_j. \tag{2.40}$$

The EDM depends on the choice of norm via the angular measure. However, Proposition 3 in Larsson and Resnick (2012) states that EDMs under different norms are equivalent in the sense of Definition 1 in the same paper. The original definition of the EDM, restricted to the bivariate case $\boldsymbol{X} = (X_1, X_2)$, instead used

$$f(\boldsymbol{\theta}) = \left(\frac{4}{\pi}\right)^2 \arctan\left(\frac{\theta_2}{\theta_1}\right) \left[\frac{\pi}{2} - \arctan\left(\frac{\theta_2}{\theta_1}\right)\right]. \tag{2.41}$$

The original and refined versions are equivalent in the same sense.

The interpretation of the coefficient is that $\mathrm{EDM}_{ij} = 0$ if and only if X_i and X_j are asymptotically independent. This follows directly from (2.35) with $\beta = \{i, j\}$, since if $\chi_{ij} = 0$ then

$$\mathrm{EDM}_{ij} = \int_{\{\boldsymbol{\theta} \in \mathbb{S}_+^{d-1} \,:\, \theta_i, \theta_j > 0\}} \theta_i \theta_j \, \mathrm{d}H(\boldsymbol{\theta}) = 0. \tag{2.42}$$

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

2.2.7 Inference

By imposing a regularity structure on the joint distributional tail, MRV facilitates – and provides a rigorous theoretical justification for – a straightforward way of extrapolating the probability law from moderately large values to more extreme tail regions. To see this, note that from ?? we have

$$\boldsymbol{\Theta} \mid (R > t) \overset{d}{\to} H(\cdot), \qquad (t \to \infty). \tag{2.43}$$

The measure H represents the limiting distribution of the angles of high threshold exceedances. This interpretation informs the general approach underpinning multivariate extreme value statistics.

2.2.7.1 Framework and notation

Generally speaking, inference for multivariate extremes involves selecting a high threshold $t > 0$ and using the information from angular components corresponding to radial threshold exceedances. 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.

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 $\boldsymbol{X}$. Let $\|\cdot\|$ be a fixed norm on $\mathbb{R}^d$. For $i \geq 1$, denote by

$$R_i := \|\boldsymbol{X}_i\| > 0, \qquad \boldsymbol{\Theta}_i := (\Theta_{i1}, \ldots, \Theta_{id}) = \frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|} \in \mathbb{S}_+^{d-1}, \tag{2.44}$$

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

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

where $\boldsymbol{X}_{(i),n} := \boldsymbol{X}_{\sigma(i)}$ for $i = 1, \ldots, n$. We call $\|\boldsymbol{X}_{(j),n}\|$ 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. For $i = 1, \dots, n$, the radial and angular components of $\boldsymbol{X}_{(i)}$ shall be denoted by

$$R_{(i)} = \|\boldsymbol{X}_{(i)}\| > 0, \qquad \boldsymbol{\Theta}_{(i)} = (\Theta_{(i),1}, \dots, \Theta_{(i),d}) = \frac{\boldsymbol{X}_{(i)}}{\|\boldsymbol{X}_{(i)}\|} \in \mathbb{S}_+^{d-1}. \tag{2.45}$$

Inference based on the $k = k(n)$ largest observations is equivalent to setting the radial threshold as $t = \hat{t}_k := R_{(k+1)}$. Only the angles $\boldsymbol{\Theta}_{(1)}, \dots, \boldsymbol{\Theta}_{(k)}$ will enter into the estimators.

In theoretical analyses, it is customary to choose the sequence $\{k(n) : n \geq 1\}$ such that

$$\lim_{n\to\infty} k(n) = \infty, \qquad \lim_{n\to\infty} \frac{k(n)}{n} = 0. \tag{2.46}$$

These arise as sufficient conditions for proving asymptotic properties (e.g. consistency) of estimators. The first condition ensures that the effective sample size becomes arbitrarily large. The second condition means that the proportion of threshold exceedances becomes vanishingly small, so that the estimators focus further into the tail. The more challenging practical question of how to select k will be discussed later in Section XXX.

2.2.7.2 The empirical angular measure

The empirical angular measure is the natural non-parametric estimator for the angular measure. It represents the empirical distribution of the angles of the set of threshold exceedances.

Definition 2.13. The empirical angular measure based on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ is the random measure on $\mathbb{S}_+^{d-1}$ defined as

$$\hat{H}(\cdot) := \frac{m}{k} \sum_{i=1}^{n} \delta_{\boldsymbol{\Theta}_i}(\cdot) \mathbf{1}\{R_i > \hat{t}_k\} = \frac{m}{k} \sum_{i=1}^{k} \delta_{\boldsymbol{\Theta}_{(i)}}(\cdot). \tag{2.47}$$

Note that $\hat{H}$ does not enforce the moment constraints (2.14), so is not necessarily a valid angular measure. Einmahl and Segers (2009) propose 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 (2.46). 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 result holds with $\|\cdot\| = \|\cdot\|_p$ for $p \in [1,\infty]$. The empirical angular measure is a discrete angular measure on k points. Consequently, there exists a max-linear random vector with k factors whose angular measure is $\hat{H}$ (ignoring the fact that $\hat{H}$ is not necessarily a valid angular measure).

The empirical angular measure is used to construct further non-parametric estimators. One is often interested in quantities of the form

$$\mathbb{E}_{\boldsymbol{\Theta}\sim H}[f(\boldsymbol{\Theta})] := \int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta})\,\mathrm{d}H(\boldsymbol{\theta}), \tag{2.48}$$

where $f : \mathbb{S}_+^{d-1} \to \mathbb{R}$. Unlike Klüppelberg and Krali (2021), our notation retains the mass of M in H (rather than absorbing it into f), so that if $\tilde{H} = m^{-1}H$ denotes the normalised counterpart of H, then

$$\mathbb{E}_{\boldsymbol{\Theta}\sim H}[f(\boldsymbol{\Theta})] = \int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta})\,\mathrm{d}H(\boldsymbol{\theta}) = m\int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta})\,\mathrm{d}\tilde{H}(\boldsymbol{\theta}) = m\mathbb{E}_{\boldsymbol{\Theta}\sim\tilde{H}}[f(\boldsymbol{\Theta})].$$

The analogous relation for variances is

$$\mathrm{Var}_{\boldsymbol{\Theta}\sim H}(f(\boldsymbol{\Theta})) = m^2\mathrm{Var}_{\boldsymbol{\Theta}\sim\tilde{H}}(f(\boldsymbol{\Theta})).$$

A natural estimator of (2.48) is obtained by replacing H with the discrete random measure $\hat{H}$ in the right-hand side, yielding

$$\hat{\mathbb{E}}_{\boldsymbol{\Theta}\sim H}[f(\boldsymbol{\Theta})] := \mathbb{E}_{\boldsymbol{\Theta}\sim\hat{H}}[f(\boldsymbol{\Theta})] = \int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta})\,\mathrm{d}\hat{H}(\boldsymbol{\theta}) = \frac{m}{k}\sum_{i=1}^{k} f(\boldsymbol{\Theta}_{(i)}). \tag{2.49}$$

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*

(2.46). *Moreover, suppose that*

$$\lim_{n\to\infty} \sqrt{k}\left[\frac{n}{k}\mathbb{E}[f(\boldsymbol{\Theta}_1)\mathbf{1}\{R_1 \geq b_{\lfloor n/k\rfloor} t^{-1/\alpha}\}] - \mathbb{E}_{\boldsymbol{\Theta}\sim H}[f(\boldsymbol{\Theta})]\frac{n}{k}\bar{F}_R(b_{\lfloor n/k\rfloor} t^{-1/\alpha})\right] = 0 \qquad (2.50)$$

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*

$$\sigma^2 := \mathrm{Var}_{\boldsymbol{\Theta}\sim H}(f(\boldsymbol{\Theta})) > 0. \qquad (2.51)$$

Then

$$\sqrt{k}\left[\hat{\mathbb{E}}_{\boldsymbol{\Theta}\sim H}[f(\boldsymbol{\Theta})] - \mathbb{E}_{\boldsymbol{\Theta}\sim H}[f(\boldsymbol{\Theta})]\right] \to N(0,\sigma^2), \qquad (n\to\infty). \qquad (2.52)$$

The rate condition (2.50) requires that the dependence between the radius and angle decays sufficiently quickly. This condition is non-observable and must be assumed. For $f(\boldsymbol{\theta}) = \theta_i\theta_j$ the condition (2.51) excludes the case of asymptotic independence, i.e. $\mathrm{EDM}_{ij} = 0$, since the limit distribution is degenerate. This prevents us from, say, establishing asymptotic normality of $\widehat{\mathrm{EDM}}_{ij} = \hat{\mathbb{E}}_{\boldsymbol{\Theta}\sim H}[\Theta_i\Theta_j]$ under asymptotic independence. In that case, the above result would only prove consistency $\widehat{\mathrm{EDM}}_{ij} \to 0$. Possible strategies for circumventing this issue are proposed in Lehtomaa and Resnick (2020).

2.3 Tail pairwise dependence matrix (TPDM)

2.3.1 Equivalent definitions

Preamble here.

Definition 2.14. 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

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} = \int_{\mathbb{S}_{+(2)}^{d-1}} \theta_i\theta_j \,\mathrm{d}H(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\Theta}\sim H}[\Theta_i\Theta_j]. \qquad (2.53)$$

This definition was generalised to permit any $\alpha \geq 1$ by Kiriliouk and Zhou (2022).

Definition 2.15. 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

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} = \int_{\mathbb{S}_{+(\alpha)}^{d-1}} \theta_i^{\alpha/2}\theta_j^{\alpha/2}\,\mathrm{d}H(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\Theta}\sim H}[\Theta_i^{\alpha/2}\Theta_j^{\alpha/2}]. \tag{2.54}$$

Note that the parameter α does not solely represent the tail index of $\boldsymbol{X}$. It also dictates the normalisation sequence and the choice of norm. TPDM theory requires that all these choices conform in this way. Clearly the two definitions above coincide when $\alpha = 2$, but Kiriliouk and Zhou (2022) provide no direct rationale for why (2.54) is the natural generalisation of (2.53). 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 2.1. *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

$$\tilde{h}_\alpha(\theta) = \alpha\theta^{\alpha-1}\tilde{h}_1(\theta^\alpha). \tag{2.55}$$

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 (2.29) 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 2.1. *Using the notation of Lemma 2.1, the off-diagonal entry in the TPDM of $\boldsymbol{X}$ is*

$$\sigma_{ij} = m\int_0^1 \sqrt{u(1-u)}\,\tilde{h}_1(u)\,\mathrm{d}\phi. \tag{2.56}$$

Proof. The relation between the normalised measure H_α and the measure H in Definition 2.15 is $H_\alpha = m^{-1}H$, where m is the mass of H. Therefore, (2.54) 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 2.1 and substitute $u = \theta^\alpha$ to obtain the final result

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

□

This means the TPDM is invariant under the choice of α. (Later we will show that the quantity m does not depend on α when the margins are pre-processed in a suitable way.) In principle we are free to leave α unspecified or set at some arbitrary value. Typically we will choose $\alpha = 2$, since much of the original theory and accompanying methods were developed in this setting. It also eases the notation by allowing us to omit the cumbersome $\alpha/2$ exponents. The exception to this is in Chapter XXX, where we will choose $\alpha = 1$. This affords us the ability to leverage statistical theory from the field of compositional data analysis, which pertains to random vectors on $\mathbb{S}_{+(1)}^{d-1}$.

2.3.2 Interpretation of the TPDM entries

Instead of defining the TPDM entry-wise, one can write it more succinctly as

$$\Sigma = \mathbb{E}_{\boldsymbol{\Theta}\sim H}\left[\boldsymbol{\Theta}^{\alpha/2}(\boldsymbol{\Theta}^{\alpha/2})^T\right], \tag{2.57}$$

Not coincidentally, this bears a striking resemblance to the definition of a covariance matrix in the non-extreme setting. Recall that the covariance matrix represents the second-order (central) moment of a random vector. Its diagonal entries correspond to the scale (variance) of the components. Its off-diagonal entries summarise the strength of association (unnormalised correlation) between pairs of variables. The TPDM entries can be interpreted analogously, except the notions of scale and association are adapted to refer to properties of the joint distributional tail.

Definition 2.16. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ with fixed normalisation sequence b_n. For $i = 1, \dots, d$, the scale of X_i is defined as (**kluppelberg_estimating_2021**)

$$\mathrm{scale}(X_i) = \left[\int_{\mathbb{S}_+^{d-1}} \theta_i^\alpha \, \mathrm{d}H(\boldsymbol{\theta}) \right]^{1/\alpha}. \tag{2.58}$$

The quantity is so called because it yields information about the scale of the marginal distributions, since

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

Moreover, it behaves as a measure of scale since for any $c > 0$,

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

The relation between the diagonal entries and the marginal scales is $\mathrm{scale}(X_i) = \sigma_{ii}^{1/\alpha}$.

Lemma 2.2. *Assume $\boldsymbol{X}$ is pre-processed to have Fréchet margins (2.30). Then*

1. *For all $i = 1, \dots, d$, $\sigma_{ii} = 1$.*
2. *The trace of the TPDM is $\mathrm{trace}(\Sigma) = d$.*
3. *The mass of the angular measure is $m = d$.*

Proof. For (i), we simply substitute the Fréchet survivor function, yielding

$$\sigma_{ii} = \text{scale}(X_i)^\alpha = \frac{\lim_{n\to\infty} n\mathbb{P}(X_i > n^{1/\alpha}x)}{x^{-\alpha}} = \frac{\lim_{n\to\infty} n\left\{1 - \exp\left[-(n^{1/\alpha}x)^{-\alpha}\right]\right\}}{x^{-\alpha}} = 1. \tag{2.59}$$

Statement (ii) is an obvious corollary of (i). For (iii), recall that the norm index p matches the tail index α and note that $\sum_{i=1}^d \theta_i^\alpha = \|\boldsymbol{\theta}\|_\alpha^\alpha = 1$ for any $\boldsymbol{\theta} \in \mathbb{S}_{+(\alpha)}^{d-1}$. It follows that

$$\text{trace}(\Sigma) = \sum_{i=1}^d \int_{\mathbb{S}_{+(\alpha)}^{d-1}} \theta_i^\alpha \,\mathrm{d}H(\boldsymbol{\theta}) = \int_{\mathbb{S}_{+(\alpha)}^{d-1}} \sum_{i=1}^d \theta_i^\alpha \,\mathrm{d}H(\boldsymbol{\theta}) = \int_{\mathbb{S}_{+(\alpha)}^{d-1}} \mathrm{d}H(\boldsymbol{\theta}) = m. \tag{2.60}$$

Combining (ii) and (2.60) completes the proof.

□

This result means that standardising to Fréchet margins is akin to working with re-scaled variables with unit variance in the non-extremes setting. The appropriate analogue then becomes the correlation rather than covariance matrix.

Comparing Definition 2.14 with Definition 2.12 reveals that the TPDM's off-diagonal entries are pairwise EDMs. Thus the interpretation of these entries is inherited from the EDM: X_i and X_j are asymptotically independent if and only $\sigma_{ij} = \sigma_{ji} = 0$; the magnitude of $\sigma_{ij} > 0$ reveals the strength of tail dependence between X_i and X_j.

In summary, the diagonal entries pertain to the marginal scales while the off-diagonals quantify the pairwise dependence strengths. This underlines the clear analogy between the TPDM and covariance matrices that we are familiar with in non-extreme settings. Throughout this thesis, we will employ 'heatmap' plots to visualise matrices, including the TPDM. An example is provided in FIGURE XXX, which depicts a Hüsler-Reiss parameter matrix and the corresponding TPDM. The method used to derive the model TPDM is explained in the following section – see Example 2.6}.

Figure with an example TPDM.

2.3.3 TPDMs under parametric models

We now compute the TPDM for a selection of parametric models. Parametric angular densities are typically specified for the $\alpha = 1$ case, i.e. with respect to standard Fréchet margins and the L_1-norm. Happily, Proposition 2.1 provides the formula for calculating the TPDM from such functions. We assume Fréchet margins as in (2.30), so that we may substitute $m = 2$ into (2.56). We reiterate that the following expressions hold for any choice of $\alpha \geq 1$. Invariably these expressions will involve intractable integrals. The angular densities are provided by *(CITE thesis entitled Inference on the Angular Distribution of Extremes.)*.

Example 2.5. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d)$ follows the symmetric logistic distribution with dependence parameter $\gamma \in (0,1)$. Then

$$\tilde{h}_1(\theta;\gamma) = \frac{1-\gamma}{2\gamma}[\theta(1-\theta)]^{\frac{1}{\gamma}-2}[\theta^{1/\gamma} + (1-\theta)^{1/\gamma}]^{\gamma-2}, \tag{2.61}$$

$$\sigma_{ij}(\gamma) = \frac{1-\gamma}{\gamma}\int_0^1 [u(1-u)]^{\frac{1}{\gamma}-\frac{3}{2}}[(1-u)^{1/\gamma} + u^{1/\gamma}]^{\gamma-2}\,\mathrm{d}u. \tag{2.62}$$

The limiting cases are $\lim_{\gamma\to 0}\sigma_{ij}(\gamma) = 1$ (full asymptotic dependence) and $\lim_{\gamma\to 1}\sigma_{ij} = 0$ (asymptotic independence).

Example 2.6. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d)$ follows the Hüsler-Reiss distribution with parameter matrix $\Lambda = (\lambda_{ij}^2)$. Then,

$$\tilde{h}_1(\theta;\lambda) = \frac{\exp(-\lambda/4)}{4\lambda[\theta(1-\theta)]^{3/2}}\phi\left(\frac{1}{2\lambda}\log\left(\frac{\theta}{1-\theta}\right)\right), \tag{2.63}$$

$$\sigma_{ij}(\Lambda) = \int_0^1 \frac{\exp(-\lambda_{ij}/4)}{2\lambda_{ij}u(1-u)}\phi\left(\frac{1}{2\lambda_{ij}}\log\left(\frac{u}{1-u}\right)\right)\mathrm{d}u. \tag{2.64}$$

The solid lines in Figure 2.2 depict σ_{ij} (blue) and χ_{ij} (red) as functions of the dependence parameter for the bivariate symmetric logistic and Hüsler-Reiss distributions. The dependence measures take different values (i.e. $\sigma_{ij} \neq \chi_{ij}$ in general) but the qualitative features of the curves are the same. In each case, the strength of association is a decreasing function of the model parameter. Perfect asymptotic dependence and asymptotic independence occur as the parameter approaches zero and its upper limit, respectively. For the Hüsler-Reiss distribution, both metrics indicate that dependence essentially vanishes beyond $\lambda \approx 3$. In order to empirically verify our analytical formulae, we overlay sample-based estimates of σ_{ij} (blue points) and χ_{ij} (red points). Each estimate is derived from $n = 5 \times 10^5$ independent samples and $\alpha = 2$. The data are generated using the `rmev` function in the `mev` package. Due to the abundance of samples, it is reasonable to neglect the influence of estimation error; this aspect will be examined in Section XXX. The empirical estimates of the tail dependence coefficient are taken as $\hat{\chi}_{ij}(0.9995)$. Estimates of σ_{ij} are derived from the empirical TPDM, to be defined later. Reassuringly, the empirical estimates closely align with the curves, corroborating our formulae.

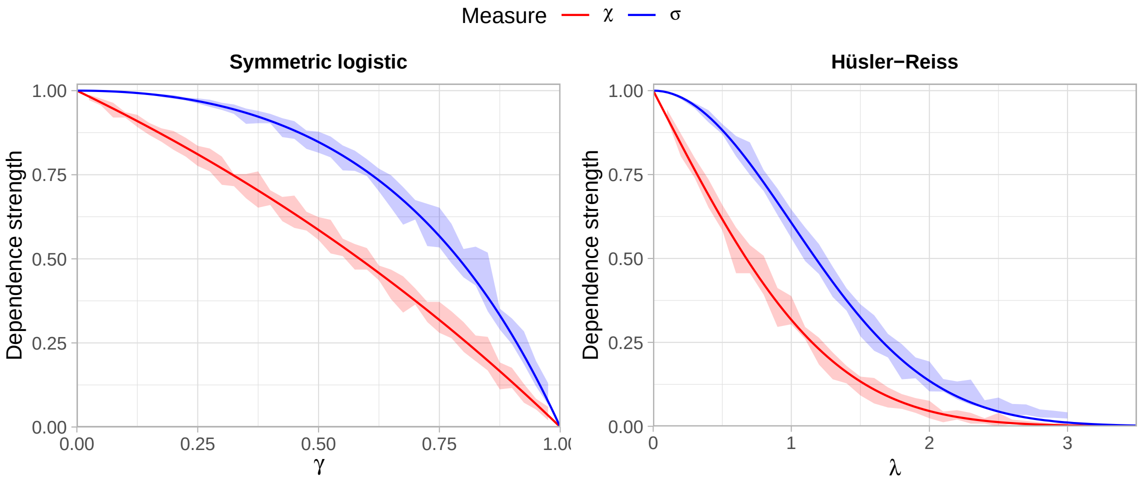

Figure 2.2: True dependence strengths for the symmetric logistic (left) and Hüsler-Reiss (right) models, measured using the tail dependence coefficient (red line) and TPDM (blue line). The shaded regions represent the minimum/maximum values of empirical estimates over 10 repeated simulations using bivariate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

Example 2.7. $\Sigma = A^{\alpha/2}(A^{\alpha/2})^T$.

Example 2.8. *To do. See density function in §3.1 of Beranger and Padoan (2015). Try a similar example to page 19 of Simpson thesis, e.g. trivariate with $\chi_{\{1,2,3\}} = 0$ to simplify the integrals. Simpson gives formulae for χ_{ij} in this case.*

2.3.4 Decompositions of the TPDM

We have established that the TPDM is useful as a summary statistic for quantifying pairwise dependencies. But one could just as easily use $\chi = (\chi_{ij})$ for this purpose, so what sets the TPDM apart? The answer lies in its additional mathematical properties. In particular, it admits two types of decomposition: eigendecomposition and the completely positive decomposition (**cooley_decompositions_2019**). These factorisations underpin most statistical applications of the TPDM, which will be reviewed in Section XXX.

Proposition 2.2. *The TPDM is symmetric and positive semi-definite (Kiriliouk and Zhou 2022, Proposition 2.1).*

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

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

Hence $\Sigma = \Sigma^T$. For any $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$. By (2.57),

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

□

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 $\boldsymbol{u}_1, \dots, \boldsymbol{u}_d \in \mathbb{R}^d$.

Definition 2.17. A matrix $M \in \mathbb{R}^{d \times d}$ is completely positive if there exists a matrix $B \in \mathbb{R}_+^{d \times q}$ such that $M = BB^T$.

Proposition 2.3. *The TPDM is completely positive. (Kiriliouk and Zhou 2022, Proposition 2.2(ii))*

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 $\boldsymbol{X}_q \sim \text{MaxLinear}(A_q, \alpha)$. For $q \geq 1$, the TPDM of $\boldsymbol{X}_q$ is $\Sigma_q = A_q^{\alpha/2} (A_q^{\alpha/2})^T$ by Example 2.7. By construction, $\{\Sigma_q : q \geq 1\}$ is a sequence of completely positive matrices. By Theorem 2.2 in CITE Berman & Shaked-Monderer (2003), the limit $\Sigma = \lim_{q \to \infty} \Sigma_q$ is also completely positive.

□

Kiriliouk and Zhou (2022) provide an iterative algorithm for constructing completely positive factorisation of an arbitrary TPDM. *Summarise the algorithm and give details about CP decomposition, e.g. estimating q.*

2.3.5 The empirical TPDM}

Definition 2.18. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ on Fréchet margins (2.30) 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 is the $d \times d$ matrix

$$\hat{\Sigma} = (\hat{\sigma}_{ij}), \qquad \hat{\sigma}_{ij} = \int_{\mathbb{S}_+^{d-1}} \theta_i^{\alpha/2} \theta_j^{\alpha/2} \, \mathrm{d}\hat{H}(\boldsymbol{\theta}) = \frac{d}{k} \sum_{l=1}^{k} \Theta_{(l),i}^{\alpha/2} \Theta_{(l),j}^{\alpha/2}. \tag{2.65}$$

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.

2.3.5.1 Finite-sample properties

Proposition 2.4. *The empirical TPDM is completely positive.*

Proof. Consider the matrix

$$\hat{A} := \left(\frac{d}{k}\right)^{1/\alpha} \left(\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}\right) \in \mathbb{R}_+^{d \times k}. \tag{2.66}$$

Note that A is $d \times k$ with non-negative entries. Then

$$\hat{A}^{\alpha/2} (\hat{A}^{\alpha/2})^T = \frac{d}{k} \sum_{i=1}^{k} \boldsymbol{\Theta}_{(i)}^{\alpha/2} \left(\boldsymbol{\Theta}_{(i)}^{\alpha/2}\right)^T = \hat{\Sigma}. \tag{2.67}$$

□

Proposition 2.5. *The empirical TPDM is symmetric and positive semi-definite.*

Proof. Let $\hat{A}$ be as in (2.66). Then for any $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$,

$$\boldsymbol{y}^T \hat{\Sigma} \boldsymbol{y} = \boldsymbol{y}^T \hat{A} \hat{A}^T \boldsymbol{y} = \|\hat{A}^T \boldsymbol{y}\|_2^2 \geq 0. \tag{2.68}$$

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

□

2.3.5.2 Asymptotic properties

Proposition 2.6. *Assume the conditions of Theorem 2.3 hold. Then the entries of $\hat{\Sigma}$ are consistent and asymptotically normal, that is, for any $i, j = 1, \ldots, d$,*

$$\sqrt{k}(\hat{\sigma}_{ij} - \sigma_{ij}) \to \mathrm{N}(0, \nu_{ij}^2), \tag{2.69}$$

where

$$\nu_{ij}^2 := \mathrm{Var}_{\boldsymbol{\Theta} \sim H}(\Theta_i^{\alpha/2} \Theta_j^{\alpha/2}).$$

Proof. Follows by application of Theorem 2.3 with the continuous function $f(\boldsymbol{\theta}) = \theta_i^{\alpha/2} \theta_j^{\alpha/2}$.

□

Adopting the notation of Proposition 2.1, 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 2.1 gives the final expression

$$\nu_{ij}^2 = m^2 \int_0^1 u(1-u)\, \tilde{h}_1(u) \,\mathrm{d}u - \left[m \int_0^1 \sqrt{u(1-u)}\, \tilde{h}_1(u) \,\mathrm{d}u \right]^2. \tag{2.70}$$

The asymptotic distribution of $\hat{\sigma}_{ij}$ does not depend on α. By **??** we have that

$$\lim_{n \to \infty} \mathbb{P}\left[\hat{\sigma}_{ij} \in \left(\sigma_{ij} - z_{\beta/2} \frac{\nu_{ij}}{\sqrt{k}}, \sigma_{ij} + z_{\beta/2} \frac{\nu_{ij}}{\sqrt{k}} \right) \right] = 1 - \beta,$$

where $z_{\beta/2} = \Phi^{-1}(1-\beta/2)$. If the angular density of (X_i, X_j) is known, then the bounds of the interval can be computed and their values do not depend on α.

Suppose $\boldsymbol{X} = (X_1, \dots, X_d) \in \mathcal{RV}_+^d(\alpha)$ is symmetric logistic with dependence parameter $\gamma = 0.6$. Using the density function in Example 2.5 and the formulae (2.56) and (2.70), we obtain by numerical integration $\sigma_{ij} \approx 0.760$ and $\nu_{ij}^2 = 0.065$ for all $i \neq j$. For sufficiently large n,

$$\mathbb{P}\left[\hat{\sigma}_{ij} \in \left(0.847 \pm \frac{1.96\sqrt{0.0358}}{\sqrt{k}}\right)\right] \approx 0.95.$$

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

The following result generalises asymptotic normality of the empirical TPDM to the entire matrix, rather than just individual entries.

Proposition 2.7. *$\hat{\Sigma}$ possesses consistency and asymptotically normality. By this, we mean that the upper-half vectorised empirical TPDM*

$$\hat{\boldsymbol{\sigma}} := \mathrm{vecu}(\hat{\Sigma}) := (\hat{\sigma_{12}}, \hat{\sigma}_{13}, \dots, \hat{\sigma}_{1d}, \hat{\sigma}_{23}, \dots, \hat{\sigma}_{2d}, \dots, \hat{\sigma}_{d-1,d})$$

is asymptotically multivariate normal,

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

where $\boldsymbol{\sigma} := \mathrm{vecu}(\Sigma)$ is defined analogously to $\hat{\boldsymbol{\sigma}}$. The diagonal and off-diagonal entries of the $\binom{d}{2} \times \binom{d}{2}$ asymptotic covariance matrix $V = (v_{ij,lm})$ are given by

$$v_{ij,lm} := \lim_{k\to\infty} k\mathrm{Cov}(\hat{\sigma}_{ij}, \hat{\sigma}_{lm}) = \begin{cases} \nu_{ij}^2, & (i,j) = (l,m), \\ \rho_{ij,lm} & \textit{otherwise}, \end{cases} \tag{2.71}$$

where

$$\rho_{ij,lm} := \frac{1}{2}\left[\mathrm{Var}_{\boldsymbol{\Theta}\sim H}(\Theta_i^{\alpha/2}\Theta_j^{\alpha/2} + \Theta_l^{\alpha/2}\Theta_m^{\alpha/2}) - \nu_{ij}^2 - \nu_{lm}^2\right]. \tag{2.72}$$

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 2.6. The off-diagonal entries are given by

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

□

Note that in the vectorisation step we only include the strictly upper triangular elements of the TPDM. One could include the diagonal entries and the result still holds, but the limiting distribution would be degenerate. The reason for this is that the diagonal TPDM entries sum to d, so V would be singular. The following example illustrates a rare case where it is possible to compute the exact asymptotic distribution of $\mathrm{vecu}(\hat{\Sigma})$.

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 $i,j=1,\dots,d$, 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

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

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

$$v_{ij,lm} = d\sum_{s=1}^{q}\frac{(a_{is}a_{js}a_{ls}a_{ms})^{\alpha/2}}{\|\boldsymbol{a}_s\|_\alpha^\alpha} - \sigma_{ij}\sigma_{lm}. \tag{2.73}$$

Suppose A is as shown in Figure XXX (left). This matrix has $q=8$ columns and $d=4$ rows summing to unity. The TPDM of $\boldsymbol{X}=A\times_{\max}\boldsymbol{Z}$ (see (2.23)) is displayed in the middle plot. The right-hand plot shows the asymptotic covariance matrix V, calculated using (2.73). The number of rows/columns in V is $\binom{4}{2}=6$. Figure XXX shows pairwise plots of $(\hat{\sigma}_{ij},\hat{\sigma}_{lm})$ derived from 1000 samples of size $n=10^4$ with $k=\sqrt{n}$. First, consider the diagonal sub-panels. These depict the empirical (red histogram) and asymptotic distributions (blue curves) of $\hat{\sigma}_{ij}$. Specifically, each blue curve represents the density function of $\mathrm{N}(\sigma_{ij},\nu_{ij}^2/k)$ random variable. The distributions are a close match. We conclude that n is sufficiently large for the asymptotic approximation suggested by Proposition 2.6 to hold. Now we exaine the numerical values printed in the upper triangular panels. The blue numbers are the true entries $v_{ij,lm}$ of V. The red numbers are sample-based estimates $\hat{v}_{ij,lm}$ of $v_{ij,lm}$, i.e. the sample covariance of $\hat{\sigma}_{ij}$ and $\hat{\sigma}_{lm}$, multiplied by k. For all pairs these values show good agreement. Finally, consider the scatter plots in the lower triangular portion of the plot. The grey points represent realisations of $(\hat{\sigma}_{ij}, \hat{\sigma}_{lm})$ over the 1000 simulations. By Proposition 2.7, for n sufficiently large,

$$\begin{pmatrix} \hat{\sigma}_{ij} \\ \hat{\sigma}_{lm} \end{pmatrix} \dot{\sim} \mathrm{N}\left(\begin{pmatrix} \sigma_{ij} \\ \sigma_{lm} \end{pmatrix}, \frac{1}{k} \begin{pmatrix} \nu_{ij}^2 & \rho_{ij,lm} \\ \rho_{ij,lm} & \nu_{lm}^2 \end{pmatrix} \right).$$

The blue ellipses are the true 95% confidence ellipses centred at the true TPDM values (blue crosses). The angle of the ellipse 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 variances ν_{ij}^2, ν_{lm}^2. The red ellipses and red crosses represent the sample-based 95% confidence region and sample mean, respectively. *Conclusions and comments.*

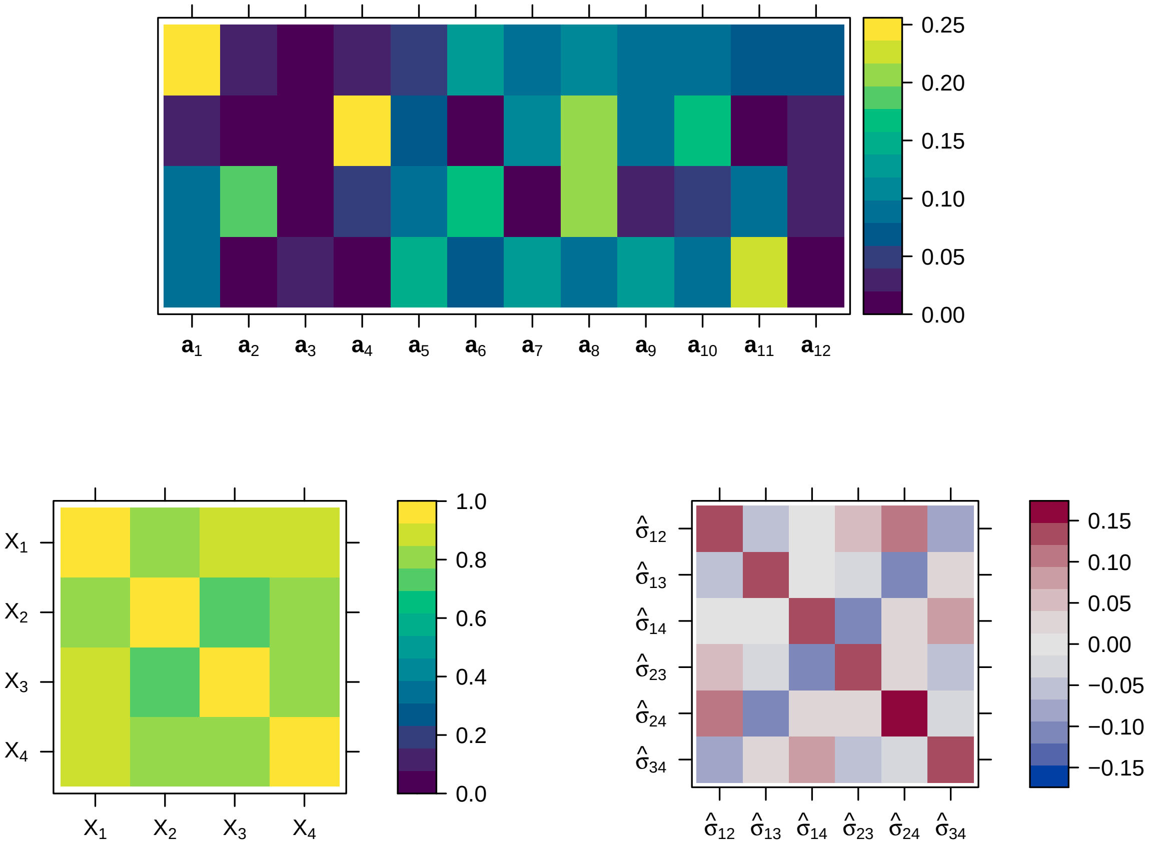

Figure 2.3: A randomly generated max-linear parameter matrix A with $d = 4$ and $q = 12$ (top), the corresponding TPDM Σ (bottom left), and the asymptotic covariance matrix V of the empirical TPDM (bottom right).

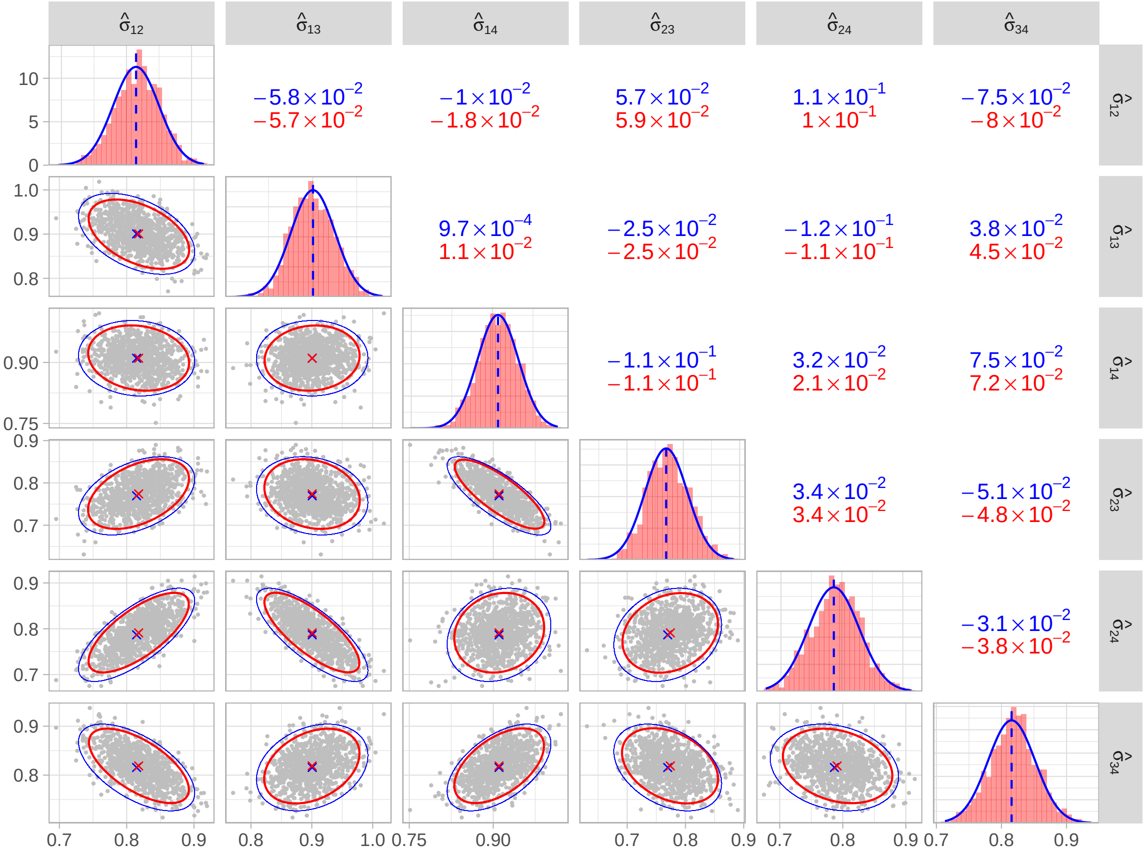

Figure 2.4: Pairs plot illustrating asymptotic normality of the empirical TPDM. Based on 1,000 empirical TPDMs estimated from samples with $n = 10^4$ and $k = 100$. The data are generated from a max-linear model with parameter matrix A as in Figure 2.3. All panels: red represents the empirical quantity based on the 1,000 repeated simulations; blue represents the theoretical limit 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.

2.4 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

(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).

2.4.1 Principal component analysis (PCA) for extremes

Definition 2.19. 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.

2.4.1.1 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

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

$\langle \cdot, \cdot \rangle_{\mathcal{H}}$. 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 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

$$R(\mathcal{S}) := \mathbb{E}[\|\boldsymbol{Y} - \Pi_{\mathcal{S}} \boldsymbol{Y}\|_2^2] \tag{2.74}$$

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 (2.74) 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 (2.74) cannot be minimised directly. Instead we resort to an empirical risk minimisation (ERM) approach, whereby the risk is replaced by

$$\hat{R}(\mathcal{S}) := \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{Y}_i - \Pi_{\mathcal{S}} \boldsymbol{Y}_i\|_2^2 \tag{2.75}$$

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

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

where $\hat{V}_{ij} := \langle \boldsymbol{Y}_i, \boldsymbol{u}_j \rangle$. The subspace $\hat{\mathcal{S}}_p = \mathrm{span}\{\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_p\}$ minimises (2.75) 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.

2.4.1.2 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 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. 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 $j = 1, \ldots, d$. Then $\mathcal{S}_p = \operatorname{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d\}$ 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 (2.46), 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 fashion may be unrealistic, since data in the simplex are prone to exhibit curvature Aitchison (1983).

2.4.1.3 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

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

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.

2.4.1.4 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 (2.65) 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 $\boldsymbol{X} = (X_1, \ldots, X_p) \in \text{RV}_+^p(2)$

and $\boldsymbol{Y} = (Y_1, \dots, Y_q) \in \mathrm{RV}_+^q(2)$, 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 $(\boldsymbol{X}, \boldsymbol{Y}) = (X_1, \dots, X_p, Y_1, \dots, Y_q) \in \mathrm{RV}_+^{p+q}(2)$ and the variable of integration is indexed as $\boldsymbol{\theta} = (\theta_1^X, \dots, \theta_p^X, \theta_1^Y, \dots, \theta_q^Y)$. (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?

2.4.2 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 $\boldsymbol{w}_1, \dots \boldsymbol{w}_K$ minimising some objective function of the form

$$\mathbb{E}_{\boldsymbol{\Theta}\sim H}\left[\min_{l=1,\dots,K} \varrho(\boldsymbol{\Theta}, \boldsymbol{w}_l)\right], \tag{2.76}$$

  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 $C_{\beta_1}, \dots, C_{\beta_K}$, where $\beta_1, \dots, \beta_K \in \mathcal{P}(\{1, \dots, d\}) \setminus \emptyset$ 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 $\beta_1, \dots, \beta_K$ may be simultaneously extreme. Identification of the support of the angular measure is notoriously challenging because the extremal angles $\boldsymbol{\Theta}_{(1)}, \dots, \boldsymbol{\Theta}_{(k)}$ 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 $X_1, \dots, X_d$ 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 $\beta_1, \dots, \beta_K$ of $\{1, \dots, d\}$ such that the angular measure is supported on $C_{\beta_1}, \dots, C_{\beta_K}$ 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 $\boldsymbol{X}_{\beta_1}, \dots, \boldsymbol{X}_{\beta_K}$. 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.

2.4.2.1 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 (2.76) 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 $\boldsymbol{\theta}_{(1)},\dots,\boldsymbol{\theta}_{(k)} \in \mathbb{S}_{+(2)}^{d-1}$ and current centroids $\hat{\boldsymbol{w}}_1,\dots,\hat{\boldsymbol{w}}_K \in \mathbb{S}_{+(2)}^{d-1}$. A single iteration of their procedure yields new centroids $\hat{\boldsymbol{w}}_1^\star,\dots,\hat{\boldsymbol{w}}_K^\star \in \mathbb{S}_{+(2)}^{d-1}$ given by the respective principal eigenvectors of

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

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 $C_{\beta_1},\dots,C_{\beta_K}$. Thresholding the centroid vectors yields the final partition $\beta_1,\dots,\beta_K$.

2.4.2.2 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 $i = 1,\dots,d$. 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.

2.4.3 Parametric model fitting

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

$$\boldsymbol{X} = (I - \rho W)^{-1} \otimes \boldsymbol{Z}. \tag{2.77}$$

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 (2.24) 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

$$\hat{\rho} = \underset{\rho\in(0,1/4)}{\arg\min}\, \|\hat{\Sigma} - \Sigma(\rho)\|_F^2. \tag{2.78}$$

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 (2.66) 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.*

2.4.4 Miscellaneous: time series and extremal graphical models

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

2.5 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 (2.46). 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.

2.5.1 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

$$\sigma_{ij} \ll 1 \implies \text{Bias}(\hat{\sigma}_{ij}) = \mathbb{E}[\hat{\sigma}_{ij}] - \sigma_{ij} > 0. \tag{2.79}$$

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

2.5.2 Simulation experiments

See Figure 2.5.

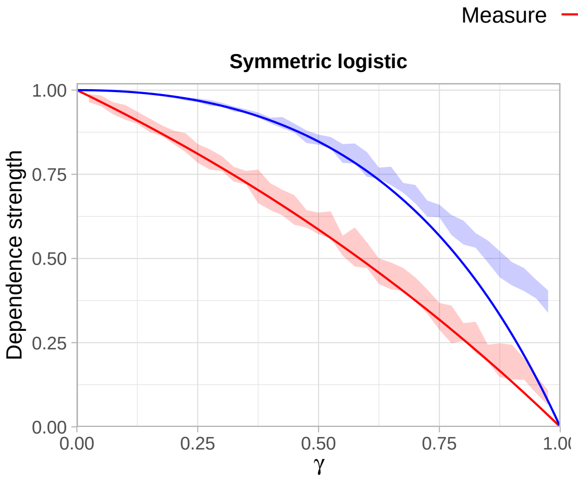

Figure 2.5: True dependence strengths for the symmetric logistic (left) and Hüsler-Reiss (right) models, measured using the tail dependence coefficient (red line) and TPDM (blue line). The shaded regions represent the minimum/maximum values of empirical estimates over 10 repeated simulations using bivariate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

2.5.3 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 (2.79) 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 (2.77). When the spatial extent of the study domain is large compared to that of the modelled phenomenon, their estimation procedure (2.78) 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

$$\tilde{\Sigma} = (\tilde{\sigma}_{ij}), \qquad \tilde{\sigma}_{ij} = \begin{cases} \hat{\sigma}_{ij}, & i = j, \\ (\hat{\sigma}_{ij} - \hat{\beta}_2)_+, & i \neq j. \end{cases} \tag{2.80}$$

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 (2.80) 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.

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.

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

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. 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.

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.

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.

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 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 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.

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.

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

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.

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.

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.